National University of Asunción

Polytechnic School

Categorical Multivariate Dependency for Feature Selection applied in Datamining Classification Task

Gustavo Daniel Sosa Cabrera

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy (Ph.D.) in Computer Science.

San Lorenzo - Paraguay

October, 2023

National University of Asunción

Polytechnic School

Categorical Multivariate Dependency for Feature Selection applied in Datamining Classification Task

Gustavo Daniel Sosa Cabrera

Advisors:

Prof. Christian E. Schaerer Serra, D.Sc.

Prof. Miguel García Torres, D.Sc.

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy (Ph.D.) in Computer Science.

San Lorenzo - Paraguay

October, 2023

Sosa Cabrera, Gustavo Daniel.

Categorical Multivariate Dependency for Feature Selection applied in Datamining Classification Task / Gustavo Sosa-Cabrera Advisors: Christian E. Schaerer S., Miguel García-Torres - - San Lorenzo : FPUNA, 2023.
i-xix; 125p; il:30cm.

Incluye Bibliografía y Anexos

Tesis Doctoral (Doctorado en Ciencias de la Computación). – UNA. Facultad Politécnica, 2023.

1. Filter feature selection. 2. Feature intercooperation. 3. Feature grouping. 4. Information theory-based measures. 5. Consistency-based measures. 6. Class-separability strategy.

SCDD 005.8
Sos715c

Hoja de Aprobación de Tesis

CATEGORICAL MULTIVARIATE DEPENDENCY FOR FEATURE SELECTION APPLIED IN DATAMINING CLASSIFICATION TASK

Gustavo Daniel Sosa Cabrera

Tesis de Doctorado aprobada el 20 de octubre de 2023 por los siguientes miembros del Jurado de Defensa:

Dr. David Becerra, Univ. de Loyola, España
Dr. Marcelo Castier, Univ. Texas A&M University at Qatar, Qatar
Dr. Vit Bubak, Universidad Paraguayo Alemana
Dr. Alejandro Giangreco, FIUNA
Dr. Andreas Ries, FPUNA
Dra. Cynthia Villalba, FPUNA
Dr. Juan Carlos Cabral, FPUNA
Dra. Rocio Botta Solano Lopez, FPUNA
Dr. Miguel Garcia Torres, Universidad Pablo Olavide, España, co-orientador de tesis
Dr. Christian E. Schaerer, FPUNA, orientador de tesis.

Prof. Dr. Horacio A. Legal Ayala
Coordinador Académico
Postgrado en Ciencias de la Computación
Facultad Politécnica
Universidad Nacional de Asunción

Prof. Dr. Christian E. Schaerer S.
Orientador

Dedicated to my beloved son Enzo,
my lovely wife Leslie,
my dear mother Teresa,
my appreciated brother Osvaldo,
and my deeply missed father Epifanio[†]

for their endless love, support and encouragement.

[†]Aunque no te llore, me duele; aunque no te hable, te pienso; aunque no te busque, te extraño. Siempre te querré, Papá.-

Acknowledgements

A frontier is never a place; it is a time and a way of life.
(Hal Borland)

First, I want to thank my advisors Professors Christian Schaerer and Miguel García Torres for all the support, patience, and encouragement he provided during my Ph.D. studies. They are truly brilliant professionals and above all great people.

Also, my sincere gratitude to the thesis committee members and the external examiners, many thanks for their constructive comments and suggestions on improving my work.

I would also like to thank all members of the NIDTEC-FPUNA, at the National University of Asunción for their help and comments during my work.

In life we meet many teachers and bosses, but only some leave traces. Thus, I would like to thank my great mentor Professor Juan Segovia Silvero because without his teachings, I would not have been able to get this far in my professional life and Professor María Elena García for her enormous quality of person and her invaluable help in the most difficult moments.

Finally, I must express my very profound gratitude to my wife, my mother, my brother, and my late father *(Who sadly passed away during my studies)* for providing me with unfailing support and continuous encouragement throughout my years of study. This accomplishment would not have been possible without them. Thank you.

Gustavo Sosa Cabrera.-

Categorical Multivariate Dependency for Feature Selection applied in Datamining Classification Task

Author: Gustavo Sosa Cabrera

Advisors:

Prof. Christian E. Schaerer Serra, D.Sc.

Prof. Miguel García Torres, D.Sc.

Summary

Nowadays, the technological progress of a constantly evolving world is related to the unprecedented increase of collected data, with hundreds or thousands of features and instances. As a result, feature selection has become an inseparable part of any preprocessing for dimensionality reduction in machine learning. However, in this Big Data era characterized by complex and heterogeneous datasets, most of the proposed feature selection methods have focused on using a single feature evaluation measure on a unique search space. In this study, a novel filter framework based on partition and intercooperation (PART_FS) is proposed. In this approach, the search space is partitioned into subspaces according to the type of information contained in the feature (i.e. individual informative, synergistic and complementary). For the analysis of redundancy, the strategy based on class separability in conjunction with Markov blanket property is used. Synergy is evaluated using an information theory measure, while complementarity is evaluated using a consistency-based measure. To show the performance of PART_FS, it was compared with five efficient cooperativeness-based feature selection methods, FS_RRC, IIFS, FJMI, SAFE and RELAX_MRMR, on three artificial data and twenty public datasets in combinations with seven classifiers. Experiment results on both artificial and real world data demonstrate the superiority of PART_FS when applied to a variety of problems with distinct characteristics. Hence, the partitioning of the search space and the differential treatment on each subspace could better assess the importance of the features in the preprocessing of complex high-dimensional data.

Keywords: Filter feature selection · Feature intercooperation · Feature grouping · Information theory-based measures · Consistency-based measures · Class-separability strategy.

Dependencia Multivariante Categórica para la Selección de Atributos aplicado en la Tarea de Clasificación de la Minería de Datos

Autor: Gustavo Sosa Cabrera

Orientadores:

Prof. Christian E. Schaerer Serra, D.Sc.

Prof. Miguel García Torres, D.Sc.

Resumen

A día de hoy, el progreso tecnológico de un mundo en constante evolución está relacionado con el aumento sin precedentes de los datos recopilados, con cientos o miles de atributos e instancias. Como resultado, la selección de atributos, se ha convertido en una parte inseparable de cualquier preprocesamiento para la reducción de la dimensionalidad en el aprendizaje de máquinas. Empero, en esta era del Big Data caracterizado por complejos y heterogéneos conjuntos de datos, la mayoría de los métodos de Selección de Atributos propuestos se han centrado en utilizar una única medida de evaluación de atributos sobre un único espacio de búsqueda. En este enfoque, el espacio de búsqueda es particionado en subespacios de acuerdo al tipo de información contenida en el atributo (*i.e.* informativo individual, sinérgico y complementario). Para el análisis de la redundancia se emplea la estrategia basada en la separabilidad de clases junto con la propiedad de Markov blanket. La sinergia es evaluada mediante una medida de la teoría de la información, mientras que la complementariedad mediante una medida basada en consistencia. Para demostrar el rendimiento de PART_FS, el mismo fue comparado con 5 eficientes métodos de Selección de Atributos basados en intercooperación, FS_RRC, IIFS, FJMI, SAFE y RELAX_MRMR, sobre tres conjuntos de datos artificiales y una veintena de conjuntos de datos del repositorio UCI en combinación con siete clasificadores. Los resultados experimentales tanto sobre los datos artificiales como también los del mundo real, demuestran la superioridad de PART_FS cuando es aplicado sobre una variedad de problemas con características diferentes. Por tanto, la partición del espacio de búsqueda y el tratamiento diferenciado sobre cada subespacio podría evaluar mejor la importancia de las características en el preprocesamiento de datos complejos de alta dimensionalidad.

Palabras clave: Filtro de Selección de Atributos · Intercooperation de Atributos · Agrupamiento de Atributos · Medidas basadas en la Teoría de la Información · Medidas basadas en Consistencia · Estrategia de Separación de Clases.

Contents

List of Tables

List of Figures

List of Algorithms

Nomenclature

BIFS *Binary Interaction based Feature Selection.*

BN *Binary Consistency Measure.*

BN-K2 *Bayes Net with K2 for Searching Network Structures.*

CFS *Correlation-Based Feature Selection.*

CIFE *Conditional Informative Feature Extraction.*

CMICOT *Conditional Mutual Information with Complementary and Opposing Teams.*

CMIFS *Conditional Mutual Information Feature Selection.*

CMIM-2 *Conditional Mutual Information Maximization Version 2.*

CS *Class-Separability Strategy.*

CWC *Combination of Weakest Components.*

DFL *Discrete Function Learning.*

DISR *Double Input Symmetrical Relevance.*

DKL *Kullback-Leibler Divergence.*

DL *Deep Learning.*

DSPLUSMII *DSplus Multidimensional Interaction Information.*

FGMMI *Feature Grouping based on Multivariate Mutual Information.*

FIM *Feature Interaction Maximisation.*

FJMI *Five-way Joint Mutual Information.*

FRFS *FOIL Rule based Feature Subset Selection.*

FS *Feature Selection.*

FS-RRC *Feature Selection based on Relevance, Redundancy and Complementarity.*

GLOBALFS *Global Feature Selection.*

GT *Game Theory.*

ICAP/IC *Interaction Capture.*

IG *Information Gain.*

IGFS *Interaction Gain for Feature Selection.*

II *Interaction information.*

IIFS *Interaction Information Feature Selection.*

IM *Interaction Mining.*

IMFS-FD *Interaction-based Feature Selection using Factorial Design.*

IWFS *Interaction Weight based Feature Selection.*

J48 *Decision Tree based Classification.*

JMI *Joint Mutual Information.*

JMIM *Joint Mutual Information Maximisation.*

KM *Kononenko's Method.*

KNN *K-Nearest Neighborhood.*

LCC *Linear Consistency-Constrained.*

MI *Mutual information.*

MIMR *Min-Interaction Max-Relevance.*

ML *Machine Learning.*

MMI *Multivariate Mutual Information.*

MMIMRSC *Maintaining Mutual Information and Minimizing Redundancy-Synergy Coefficient.*

MRMC *Maximum Relevancy Maximum Complementary.*

MRMI *Max-Relevance Max-Interaction.*

MSU *Multivariate Symmetrical Uncertainty.*

NB *Naive Bayes.*

NIWFS *Neighborhood Interaction Weight based Feature Selection.*

OFS-MI *Optimal Feature Selection using Mutual Information.*

PART-FS *Feature Selection based on Partition and Intercooperation.*

PRBC *Part Rules Based Classifier.*

RCDFS *Redundancy-Complementariness Dispersion Feature Selection.*

RELAXMRMR *Relax Maximum Relevance Minimal Redundance.*

RF *Random Forest.*

SAFE *Self-Adaptive Feature Evaluation.*

SBE *Sequential Backward Elimination.*

SFS *Sequential Forward Selection.*

SU *Symmetrical Uncertainty.*

SV *Shapley Value.*

SVM *Support Vector Machine.*

TC *Total Correlation.*

UCI *University of California Irvine Machine Learning Repository.*

Chapter 1

Introduction

In classification tasks, a feature (i.e., an independent variable) is considered relevant, irrelevant, or redundant based on the information it contains about the target concept or class (i.e., the dependent variable).

Feature selection is defined as the method of finding a minimum set of relevant features in order to minimize the error in the classification process with respect to a given class.

In this regard, feature selection has become the focus of much of the research in areas involving high-dimensional datasets. Among these areas are text processing, gene expression and combinatorial chemistry (Sui, 2013).

A feature selection method has three components: the definition of the evaluation criterion (e.g., the relevance of the features), the estimation of the evaluation criterion (i.e., the measure) and the search strategies for generating subsets of candidate features.

Regarding the evaluation measures, several criteria have been proposed to evaluate the features and determine their importance. It should be noted that based on the evaluation criteria, the feature selection methods can be divided into wrapper, filter and embedded.

In filter-type methods, the evaluation of the subset of features is carried out by assessing the intrinsic properties of the data, such as distance, consistency, entropy and correlation.

This strategy does not consider any relationship with the learning algorithm, so they are much more efficient in terms of computational resources since they are executed as a previous stage called preprocessing.

Although the literature offers a wide and varied range of filter-type attribute selection methods, in most cases, it deals only with the identification of irrelevant and redundant attributes, where an important aspect that is usually neglected is the com-

plementarity of the attributes (Guyon and Elisseeff, 2003; Chen et al., 2015) (also known as (Zeng et al., 2015a) synergy or (Jakulin and Bratko, 2003a) interaction).

Interacting attributes are those that appear to be irrelevant or little relevant to the class when considered individually, but when combined with other attributes, can have a high correlation with the (Zeng et al., 2015b) class.

One motivation for the development of this thesis is that the interaction of attributes has received considerable attention in recent times and is attracting more and more attention from researchers (Zeng et al., 2015b,a).

Reasons is that over the decades, attribute selection methods have evolved from simple univariate relevance ranking algorithms, through relevance-redundancy trade-offs to more sophisticated approaches based on multivariate dependencies in recent years.

This tendency to capture multivariate dependence aims to obtain unique information about the class from what is defined in this study as inter-cooperation between attributes.

Therefore, the aim of this thesis is to propose ways to detect, measure and identify which associations between attributes collectively provide unique information about the explained variable or case class and their implications in the search for a minimum subset of relevant attributes.

1.1 Research Problems and Objectives

Feature selection remains and will continue to be an active field that is incessantly rejuvenating itself to answer new challenges (Liu et al., 2010).

This doctoral thesis presents new discoveries in the field of feature selection through a novel proposed method based on feature search space reduction and intercooperative feature processing. In particular, the following problems are investigated:

- How to define and examine a multivariate dependency measure based on information theory.
- How to develop a systematic summary and comparison studies to facilitate research and application of feature selection techniques based on multivariate dependencies.
- How to design a new feature selection method based on the use and exploitation of multivariate dependencies.

1.1.1 Research objetive

Examine the categorical multivariate dependence through its detection, quantification and characterization oriented to the feature selection process applied in the data mining classification task.

1.1.2 Specific objetives

- Define and explore a measure of multivariate dependence based on information theory such as Multivariate Symmetrical Uncertainty (MSU).
- Determine the limits of multivariate information measures in the most commonly used search strategies in the Feature Selection process.
- Establish and characterize the notions concerning multivariate dependence in the context of Feature Selection.
- Develop a systematic review of the state-of-the-art on feature selection heuristics based on multivariate dependence detection and/or quantification.
- Devise a heuristic for attribute selection by taking advantage of multivariate dependency detection.

1.2 Contributions

In this thesis work, different key aspects related to multivariate dependence in the context of feature selection have been addressed. The most significant contributions of this thesis are presented below:

- Definition and analysis of Multivariate Symmetrical Uncertainty (MSU) as a measure of higher order information applicable to the Feature Selection process.
- Study of MSU performance under data densities with known patterns in practice and with real-world datasets.
- Generic formulation and characterization of the notions regarding the feautre selection assisted by intercooperation.
- A state-of-the-art review of feature selection heuristics based on multivariate dependence summarizes the contributions of the different approaches found in the literature. In addition, current problems and challenges are presented in

order to identify the most promising methods given the specific knowledge gaps in the area.

- Proposal of a novel feature selection method based on feature search space partitioning and feature intercooperation. This method uses KMedoids for partitioning into subspaces, as well as using information-based and consistency-based measures to deal with inter-cooperative features.

- Development of a toolbox implemented in PYTHON to perform feature selection using multivariate dependency measures. The toolbox implements the main methods based on feature intercooperation, in addition to the proposed method.

1.3 Thesis Organization

The core parts and chapters of this thesis are derived from articles published or submitted during the doctoral research. The remainder of this is organized as follows:

- In Chapter 2, an extension of the Symmetrical Uncertainty (SU for short) measure called Multivariate Symmetrical Uncertainty (MSU for short) is proposed. This chapter is derived from (Sosa-Cabrera et al., 2019) and highlights that convey the core findings of the research are:

  - Define and explore MSU as a correlation measure for multiple nominal variables.
  - Introduce representativeness as desirable property of a sample from a nominal variable.
  - Prove that a non-representative sample under-estimates the actual value of MSU.
  - Calculate the sample size that assures representativeness at 1-α level of probability.
  - Show how MSU with its interaction detection can be applied to feature selection.

- Chapter 3 presents a comprehensive survey of the state-of-the-art work on filter feature selection methods assisted by feature inter-cooperation, which summarizes the contributions of different approaches found in the literature. This chapter

is derived from (Sosa-Cabrera et al., 2023) and highlights that convey the core findings of the research are:

  - A comprehensive survey on feature selection methods assisted by feature intercooperation is presented.
  - Twenty seven filter feature selection methods that adopt this approach reviewed.
  - Based on feature intercooperation perspective, issues and future research directions are presented.

- Chapter 4 introduces a novel feature selection method called PART_FS based on feature search space partition and feature intercooperation. This chapter is derived from a submitted paper and highlights that convey the core findings of the research are:

  - PART_FS employs KMedoids to partitioning the search space in subspaces.
  - PART_FS can deal with irrelevant, redundant, synergistic, and complementary features.
  - PART_FS apply measures both based-on information theory and consistency.
  - PART_FS outperforms five competing methods based-on intercooperation in terms of accuracy on twenty real-world datasets.

- Chapter 5 presents a preliminary summary of the main findings and discussion of future research directions.

1.4 Publications

The main chapters of this proposed doctoral dissertation are derived from the following articles published or submitted.

- **Sosa-Cabrera, G.**, Gómez-Guerrero, S., García-Torres, M., & Schaerer, C. E. (2023). *PART_FS: A feature selection method based on partitioning and intercooperation.* Status: In review.

- **Sosa-Cabrera, G.**, Gómez-Guerrero, S., García-Torres, M., & Schaerer, C. E. (2023). *Feature selection: a perspective on inter-attribute cooperation.* International Journal of Data Science and Analytics, 1-13.

- **Sosa-Cabrera, G.**, García-Torres, M., Gómez-Guerrero, S., Schaerer, C. E., & Divina, F. (2019). *A multivariate approach to the symmetrical uncertainty measure: application to feature selection problem.* Information Sciences, 494, 1-20.

Publications by the author in related research topics included in this thesis are:

- **Sosa-Cabrera, G.**, Torres, M. G., Guerrero, S. G., Schaerer, C. E., & Divina, F. (2018). *Understanding a multivariate semi-metric in the search strategies for attributes subset selection.* Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, 6(2).

- Gómez-Guerrero, S., Ortiz, I., **Sosa-Cabrera, G.**, García-Torres, M., & Schaerer, C. E. (2021). *Measuring Interactions in Categorical Datasets Using Multivariate Symmetrical Uncertainty.* Entropy, 24(1), 64.

- Gómez-Guerrero, S., **Sosa-Cabrera, G.**, García-Torres, M., Ortiz-Samudio, I., & Schaerer, C. E. (2021). *Multivariate Symmetrical Uncertainty as a measure for interaction in categorical patterned datasets.* Proceedings of the Entropy 2021: The Scientific Tool of the 21st Century session Information Theory, Probability and Statistics.

- Gómez-Guerrero, S., García-Torres, M., **Sosa-Cabrera, G.**, Sotto-Riveros, E., & Schaerer, C. E. (2021). *Classifying dengue cases using CatPCA in combination with the MSU correlation.* Proceedings of the Entropy 2021: The Scientific Tool of the 21st Century session Entropy in Multidisciplinary Applications.

Publications by the author in related research topics not included in this thesis are:

- **Sosa-Cabrera, G.**, Torres, M. G., Guerrero, S. G., Schaerer, C. E., & Divina, F. (2018). *Effect of Sample Representativeness in Multivariate Symmetrical Uncertainty for Categorical Attributes.* Proceedings of the Third Conference on Business Analytics in Finance and Industry.

- **Sosa-Cabrera, G.**, Torres, M. G., Guerrero, S. G., Schaerer, C. E. (2018). *Is it correlation or interaction?.* En III Encuentro de Investigadores de la Sociedad Científica del Paraguay.

Chapter 2

A Multivariate approach to the Symmetrical Uncertainty Measure: Application to Feature Selection Problem

In this chapter we propose an extension of the Symmetrical Uncertainty (SU) measure in order to address the multivariate case, simultaneously acquiring the capability to detect possible correlations and interactions among features. This generalization, denoted Multivariate Symmetrical Uncertainty (MSU), is based on the concepts of Total Correlation (TC) and Mutual Information (MI) extended to the multivariate case. The generalized measure accounts for the total amount of dependency within a set of variables as a single monolithic quantity. Multivariate measures are usually biased due to several factors. To overcome this problem, a mathematical expression is proposed, based on the cardinality of all features, which can be used to calculate the number of samples needed to estimate the MSU without bias at a pre-specified significance level. Theoretical and experimental results on synthetic data show that the proposed sample size expression properly controls the bias. In addition, when the MSU is applied to feature selection on synthetic and real-world data, it has the advantage of adequately capturing linear and nonlinear correlations and interactions, and it can therefore be used as a new feature subset evaluation method.

2.1 Introduction

In recent years, the huge advances in data collection and storage technologies have caused the creation of large, high-dimensional, complex and heterogeneous datasets, making the classification task more and more challenging. In general, when the di-

mensionality of a dataset increases, the complexity of the data and the number of non informative features grow as well. Thus, the higher the dimension of the data, the higher the risk of degrading the performance of the classifier.

Another important problem relies on the fact that not all the features have the same importance in terms of information for the task to be performed. It is widely accepted that according to the information that variables contain about the learning task, features can be classified as irrelevant, relevant and redundant. In a nutshell, a feature is said to be irrelevant if it contains no information about the concepts to be learned, while a relevant feature does contain information about such concepts. Also, a feature is considered redundant if the information it provides about the concepts to be learned is already included in another feature or subset of features. Therefore, irrelevant and some redundant features can be removed without degrading the learning task (Hall, 1998).

Feature selection techniques have been successfully used in a wide range of application areas, such as spam filtering (Méndez et al., 2019), recommender systems (Bag et al., 2019a), consumer's purchase intention (Bag et al., 2019b), etc. It is also noteworthy that feature selection faces new challenges, for example in data streaming (Palma-Mendoza et al., 2018), big data (Palma-Mendoza et al., 2018), and multi-objective approach (Kashef and Nezamabadi-pour, 2019).

The concept of *entropy* (Shannon, 1948) can be used as an indicator of uncertainty in the data. In (Singh et al., 2017), an entropy-based method was proposed to compute the weight of a given fuzzy formal concept. In fact, in multivariate data, the uncertainty associated to the information contained in such data can be addressed by means of either information granulation or multivariate approaches. In the first case, information granules correspond to some level of abstraction of the data and, therefore, each granule should be described by a specific property or concept (Höeppner and Klawann, 2008). A method based on Galois connection is proposed in (Singh, 2018) to build the m-polar fuzzy graph concept lattice that represents uncertainty in the features. In the multivariate approach, entropy-based measures were extended to handle multivariate data directly. In this context, several works have been proposed for analysis of time-series signals. In (Ahmed et al., 2012), an empirical mode decomposition was combined with multivariate sample entropy to describe the structural complexity of multivariate signals even for non-stationary data, such as brain states measured through various channels. The multivariate multiscale sample entropy was extended for long time series in (Ni et al., 2013), so that it can reveal the complexity of multivariate biological signals.

Entropy is also used in multivariate time-independent observations where uncertainty is associated to the features (Arias-Michel et al., 2016) rather than to the data observations. The *multivariate mutual information* (MMI) measure was introduced to discover multivariate relationships in biological networks (Pham et al., 2012). The MMI was also used for quantifying shared information in a multivariate network (Ball et al., 2017). In addition, the change of MMI when adding or removing some common randomness was studied (Chan et al., 2018). The work in this chapter is framed in the scope of MMI for the discovery of dependencies between features.

In classification tasks, *feature selection* consists in finding the minimal set of relevant features in such a way that the classification error is minimized. In order to identify this optimal subset of features, several criteria have been proposed to evaluate the goodness of feature subsets. Based on the evaluation criterion, feature selection methods are divided into wrapper, filter and embedded methods. Wrapper methods evaluate the subsets of features by using the learner as a black box. This approach achieves high accuracy since interactions between feature subsets and the learning model are taken into account. Its main drawback is the computational cost, presenting a high risk of overfitting. In contrast, filter strategies assess the quality of subsets of features according to intrinsic properties of the data, such as distance, consistency, entropy and correlation. This approach does not consider the interaction with the learning algorithm; therefore, it is faster than wrapper methods but may yield lower classification accuracy. Finally, embedded methods perform the feature selection during the induction of the classifier. The difference between wrapper and embedded methods lies in the use of an intrinsic model building metric during the learning process.

Recently, many entropy-based filter strategies have been proposed, as in (Avdiyenko et al., 2015; García-Torres et al., 2016; Li et al., 2016; Shishkin et al., 2016a; Jesus et al., 2017). Most methods define heuristically a criterion based on Mutual Information (MI) to evaluate feature subsets. Some of these criteria are the joint mutual information (JMI) (Yang and Moody, 1999; Bennasar et al., 2015), the conditional infomax feature extraction (CIFE) (Guo and Nixon, 2009), and the minimum-redundancy maximum relevance (mRMR) (Peng et al., 2005). A normalized variant of MI, reducing the bias effect towards nominal features characterized by a high number of categories, was proposed in (Hall, 1998) under the name of *Symmetrical Uncertainty* (SU). The SU correlation measure detects a linear or non-linear association between two attributes. Later in (Yu and Liu, 2004), a framework that uses SU was proposed to perform an analysis of relevance and redundancy.

Despite the fact that, in general, previous works achieve competitive results on

real datasets, the use of bivariate measures ignores possible dependencies among more than two features. In order to quantify such relationships, an extension of the MI was proposed in (McGill, 1954) for three variables. The extension expresses the information shared by all variables that is not present in any subset of these variables. Although this measure was originally named *interaction information*, it is also referred to as co-information (Bell, 2003). A later work generalized this measure to n variables (Jakulin and Bratko, 2003b). Based on this, a real-world multivariate textual corpus analysis is performed in (Shalizi, 2009). The analysis identifies positive and negative interactions when certain words (features) are taken together; positive interactions provide more information about the class than using either of the words on its own. From here an improved feature selection method based on multivariate mutual information was proposed.

The *total correlation* was presented as a generalization of the MI that measures the information shared by any two or more variables (Watanabe, 1960). This measure is also referred to as multi-information (Studený and Vejnarová, 1998).

Several works have applied the multivariate approach to the feature selection problem. In (Doquire and Verleysen, 2012), the calculation of MMI was compared between a group of features by using B-spline, NN-based and kernel-based estimators. Selection of features by applying a feature grouping approach based on MMI was proposed in (Mohammadi et al., 2017). A novel measure of redundant information based on co-information was introduced in (Ince, 2017). The measure calculates the pointwise contributions to the MI which are shared unambiguously among the considered variables. Due to computational tractability, an MMI-based heuristic that considers up to 3-way feature interactions was introduced in (Singha and Shenoy, 2018). This approach of limiting the number of feature interactions had also been used in previous works (Kojadinovic, 2005; Brown, 2009). However, it was suggested in (Chen et al., 2015) that ignoring higher-order terms may lead to misidentifying redundant features as relevant due to pairwise approximation.

SU was extended to the multivariate case in (Arias-Michel et al., 2016), receiving the name of *Multivariate SU* (MSU). Since MSU is an entropy-based measure, it suffers from the same types of bias as the SU. Moreover, in this case the bias can also be exacerbated by the size of the feature subset. In fact, larger feature subset sizes require more samples so as to be able to avoid sampling bias. In order to gain more insights on MSU bias, where the mechanisms are not yet well understood, an extensive experimental evaluation is needed. This motivates us to propose an analysis on the effect that the number of features, the cardinality of the features and the sample size

may have on MSU.

In sampling, undercoverage is a source of bias. It occurs when some members of the population are inadequately represented in the sample. Although sample size and coverage are two different issues, they are related because the probability of coverage bias is high under a small sample and it decreases when the sample size grows. In order to study how the MSU is affected by the coverage bias, the concept of *total representativeness* is first introduced. This concept refers to the case in which there is no coverage bias. Then, given a fixed number of features, the concept is used to derive an expression that relates the sample size to the probability of the least likely bin in the multinomial distribution induced by concatenating the features. Here, an important achievement is that the required sample size for a desired assurance of representativeness can be easily computed, leaving as unnecessary any previously used "rules of thumb".

The novelty of this chapter can be summarized as: Multivariate Correlation is now measurable through MSU for n categorical variables, where numerical variables can also be included after suitable discretization. Furthermore, a procedure is derived to determine sample size by controlling bias in the MSU measure to a required confidence level. Also, the applicability of MSU to the feature selection problem is demonstrated.

It is worth noting that historically, correlation has only been measurable for two numerical variables and has been limited to detecting linear relationships. For nonlinear relationships and/or for analyzing more than two variables, dependencies are postulated based on fitting observed data to specified models such as the various regression and ANOVA methods. However, if none of the tested models seems adequate, there is no guarantee that correlation is absent. In this context, MSU emerges as a new tool for the detection of multivariate dependencies.

The rest of the chapter is organized as follows. Section 2.2 introduces the theoretical foundations of this work, which are extended to multivariate concepts in Section 2.3. Then, the potential sources of bias are presented in Section 2.4. The various data used in this work are described in Section 2.5. Next, the experiments and discussion are in Section 2.6. Finally conclusions and future work are described in Section 2.7.

2.2 Theoretical foundations

Let X be a categorical (discrete) random variable with possible values $\{x_1, \ldots, x_k\}$ and probability mass function $P(X)$. The entropy H of the variable X is a measure of the uncertainty in predicting the value of X, and is defined as:

$$H(X) := -\sum_i P(x_i)\log_2(P(x_i)), \tag{2.1}$$

where $P(x_i)$ is the prior probability of the value x_i of X. $H(X)$ can also be interpreted as a measure of the amount of information a discrete random variable X produces, or the variety inherent to X.

Given another discrete random variable Y, the conditional entropy $H(X|Y)$ quantifies the amount of information needed to describe the outcome of X given that the value of Y is known. It is defined as follows:

$$H(X|Y) := -\sum_j \left[P(y_j) \sum_i P(x_i|y_j)\log_2(P(x_i|y_j)) \right], \tag{2.2}$$

with $P(x_i|y_j)$ the posterior probability of a value x_i for variable X given that the value of variable Y is y_j.

Using entropy and conditional entropy, the mutual information (MI) (Quinlan, 1993) is introduced (also called information gain (IG)). MI measures the reduction in uncertainty about the value of X when the value of Y is known, and is defined as:

$$MI(X|Y) := H(X) - H(X|Y). \tag{2.3}$$

Since MI measures how much the information provided by Y makes it easier to predict the value of X, it can be used as a *measure of correlation*. It should be noted that (i) if X and Y are independent then $MI(X|Y) = 0$, and (ii) if X and Y are fully correlated then $H(X|Y) = 0$ and hence $MI(X|Y) = H(X)$.

It can be shown that $MI(X|Y)$ is symmetrical, a quite convenient property for a paired measure. On the other hand, IG tends to increase its value when the number of values of X and/or Y increases, that is, it is biased towards high cardinality features. Therefore, MI has to be normalized with the entropies of the features in order to compensate such bias. This measure, called *Symmetrical Uncertainty* (Press et al., 1988) is expressed as:

$$SU(X,Y) := 2\left[\frac{MI(X|Y)}{H(X)+H(Y)}\right]. \tag{2.4}$$

Note now that (i) if X and Y are independent then $SU(X,Y) = 0$; and (ii) if X and Y are completely correlated then $IG(X|Y) = H(X) = H(Y)$ and so $SU(X,Y) = 1$. Therefore, the SU values are restricted to the range $[0,1]$.

Since SU has only been defined for pairs of variables, it fails to detect interactions among more than two features. It might fail in this detection even when applied

successively to different pairs of features. This represents an important limitation, and to overcome this problem, the measure is extended to the multivariate case.

2.3 Multivariate approach

In order to employ information theory to assess the dependency among features from a subset, we use the concept of *total correlation*, which was first described in (McGill, 1954) and discussed in detail in (Watanabe, 1960). Given a set of random variables $X_1, \ldots, X_n$, the joint entropy of the n random variables is defined as

$$H(X_{1:n}) := -\sum_{x_1} ... \sum_{x_n} P(x_1, ..., x_n) \log_2[P(x_1, ..., x_n)] \tag{2.5}$$

The total correlation $C(X_{1:n})$ is defined by:

$$C(X_{1:n}) := \sum_{i=1}^{n} H(X_i) - H(X_{1:n}) \tag{2.6}$$

It can be noted that for the case of $n = 2$, total correlation is equivalent to the mutual information. Moreover, it is always positive and a near-zero value indicates that all the variables are independent of each other, meaning that knowing the value of one variable does not provide any information regarding the values of the other variables.

In order to generalize the symmetrical uncertainty to the multivariate case, it is desirable that it has the following properties:

- The values have to be kept in the range $[0, 1]$;
- Higher values in the measure have to correspond to higher correlation among variables. A value of 0 implies that all variables are independent while a value of 1 corresponds to a perfect correlation among variables.

The multivariate approach, called *Multivariate Symmetrical Uncertainty* (MSU), has a similar expression to (2.6):

$$\begin{aligned} MSU(X_{1:n}) &:= f(n) \left[\frac{\sum_{i=1}^{n} H(X_i) - H(X_{1:n})}{\sum_{i=1}^{n} H(X_i)} \right] \\ &= f(n) \left[\frac{C(X_{1:n})}{\sum_{i=1}^{n} H(X_i)} \right] \end{aligned} \tag{2.7}$$

where $f(n)$ corresponds to the normalization factor. In order to define such factor we introduce the following lemma.

Lemma 2.3.1. *Given the subset of random variables $X_1, \ldots, X_n$, then*

$$\frac{1}{n} \leq \frac{H(X_{1:n})}{\sum_{i=1}^{n} H(X_i)} \leq 1. \tag{2.8}$$

Proof. Using the chain rule,

$$\begin{aligned} H(X_{1:n}) &:= \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_{i-1}) \\ &= H(X_1) + \sum_{i=2}^{n} H(X_i|X_1, \ldots, X_{i-1}) \geq H(X_1). \end{aligned} \tag{2.9}$$

So, it has to be proven that

$$H(X_{1:n}) \geq H(X_i), \tag{2.10}$$

for all $i \in [1, n]$, knowing that joint entropy is agglomerative.

Using the same rule, and taking into account that $H(X) \geq H(X|X_1, \ldots, X_{i-1})$ for $i \in [i, n]$, it is obtained

$$H(X_{1:n}) \leq H(X_1) + \ldots + H(X_n). \tag{2.11}$$

Combining (2.10) and (2.11), one has

$$nH(X_{1:n}) \geq \sum_{i=1}^{n} H(X_i) \tag{2.12}$$

and

$$H(X_{1:n}) \leq \sum_{i=1}^{n} H(X_i). \tag{2.13}$$

Therefore, the inequalities (2.8) are obtained again

$$\frac{1}{n} \leq \frac{H(X_{1:n})}{\sum_{i=1}^{n} H(X_i)} \leq 1$$

which concludes the proof. □

Denoting R_n by $R_n = [H(X_{1:n})]/[\sum_{i=1}^{n} H(X_i)]$, changing the sign of the inequality and adding a value of one, then

$$0 = 1 - 1 \leq 1 - R_n \leq 1 - \frac{1}{n} = \frac{n-1}{n}. \tag{2.14}$$

On the other hand, $MSU(X_{1:n}) = f(n)[1 - R_n] \in [0, 1]$. So, it is obtained

$$0 = f(n) \cdot 0 \leq f(n)[1 - R_n] \leq f(n) \cdot \frac{n-1}{n} = 1. \tag{2.15}$$

Using the last equality, it is obtained $f(n) = \frac{n}{n-1}$ and, therefore, MSU can be defined as follows

$$MSU(X_{1:n}) := \frac{n}{n-1}\left[1 - \frac{H(X_{1:n})}{\sum_{i=1}^{n} H(X_i)}\right]. \tag{2.16}$$

Thus from (2.15), the MSU values are also in the range $[0, 1]$. The tendency to obtain larger MSU values when some or all of the features have higher cardinalities still exists, and the upcoming sections will attempt to isolate possible causes through experiments.

In contrast with linear correlation and other statistical measures of association that are oriented to numeric data, the SU and the MSU are defined for *both* discrete numeric and categorical random variables with a finite number of values. In addition, since SU and MSU depend only on the probability mass function and not on the x_i values, they are invariant under translation and scale changes applied to any numeric X_i, as long as the probability weights remain the same.

Lemma 2.3.2. *The MSU measure reduces to the SU if the number of features is $n = 2$.*

Proof. Setting $n = 2$ in equation (2.16),

$$\begin{aligned} MSU(X_{1:2}) &= \frac{2}{2-1}\left[1 - \frac{H(X_{1:2})}{H(X_1) + H(X_2)}\right] \\ &= 2\frac{H(X_1) + H(X_2) - H(X_1, X_2)}{H(X_1) + H(X_2)} \end{aligned} \tag{2.17}$$

From the chain rule in (2.9), it can be written $H(X_1, X_2) = H(X_1) + H(X_2|X_1)$, therefore

$$\begin{aligned} MSU(X_{1:2}) &= 2\frac{H(X_1) + H(X_2) - [H(X_1) + H(X_2|X_1)]}{H(X_1) + H(X_2)} \\ &= 2\frac{H(X_2) - H(X_2|X_1)}{H(X_1) + H(X_2)} \\ &= SU(X_2, X_1) = SU(X_1, X_2) \end{aligned} \tag{2.18}$$

The last results arise from equations (2.3) and (2.4) and from symmetry of the SU. Thus it is observed that the SU and the MSU are actually the same measure. □

2.4 Multivariate Cardinality and Sample Size

In order to apply MSU to finite real-world data, we study how the *curse of dimensionality* affects it. When increasing the dimensionality of the data, the volume of the search space expands exponentially and, therefore, the data become sparse. Another consequence is that such sparsity is not uniformly distributed over the search space. This may yield to some kind of bias in MSU.

Definition 2.4.1. *Given a discrete random variable (RV) X, its Univariate Cardinality, denoted by $|X|$, is the number of possible distinct labels of X.*

We extend the definition to a set of RVs as follows.

Definition 2.4.2. *Given a set of discrete random variables $X_1, \ldots, X_d$ its Multivariate Cardinality is the number of possible label combinations among all features. That is $\prod_{i=1}^{d} |X_i|$.*

Note that both univariate and multivariate cardinality types provide an indication of the *diversity of information* that discrete RVs contain. We see that for entropic measures this diversity has a parallel with the measure of spread in statistics: given a sample S from a discrete RV, the required sample size to be useful for analysis and prediction purposes is calculated as a function of the standard deviation, which is a measure of *dispersion of the numeric values* of the RV.

Standard deviation is not defined for a categorical variable, but the cardinality is an expression of its *diversity of categorical labels*. Thus, when we work with categorical variables a few analogies arise, as shown in Table 2.1, and we claim it is reasonable to expect that a relationship exists between sample size and cardinality. For the moment, in the table the claim is written as a question that will be answered in later subsections.

When working with a sample of data it is very important that it be representative to accurately reflect the entire population. If so, the sample values of the numeric RV are spread around the mean in a way that resembles the population's spread. Similarly, in the non-numeric case it is important that all or most existing categories be represented in the sample (Schouten et al., 2009; Shlomo et al., 2012), so that cardinality in that sample resembles that of the population. With this motivation the following definition is introduced.

Definition 2.4.3. *Let X be a discrete random variable with finite univariate cardinality $|X|$, and let S be a sample of m subjects of X. The sample S is said to be Totally Representative with respect to X if and only if each of its $|X|$ labels is present in the sample at least once.*

Table 2.1: Analogies between Numerical and Categorical Variables for Sample Size m

Continuous and Discrete numerical variables	*Nominal and Ordinal categorical variables*
Dispersion measures defined	Entropy measures defined
Standard deviation σ measures the dispersion of numeric values	Cardinality $\|X\|$ measures the diversity of categorical labels
m is a function of σ	Is m a function of $\|X\|$?
We wish that sample values spread around the mean in a way similar to population spread	We wish to get a sample that resembles the entropy (diversity) of the feature in the population
To resemble population spread: make sure sample is selected through correct *randomization*	To resemble population diversity: get each value of feature *represented* in the sample

We now turn to deriving important properties associated to representativeness. Later on, these properties will be used to pursue a way of ensuring representativeness through proper sample size, and inducing a more stable MSU behavior.

2.4.1 Sample histogram

Given a sample from a discrete RV, the frequencies of all its categories can be represented in a histogram. In the case of a subset of features a histogram can be used for each RV; or the grouping property (Singh and Gani, 2015) can be used to agglomerate the RVs. In this case it suffices to consider a single sample histogram associated to the concatenated values that result after applying the concatenation operation to all RVs as seen in Section 2.3. For instance, given three binary RVs X_1, X_2 and X_3, their concatenation X has $k = 2^3 = 8$ possible outcomes. Furthermore, the resulting outcomes follow a multinomial distribution with probabilities implied by the underlying X_1, X_2 and X_3; if the three features are independent then the multinomial is a "flat" density with all outcomes equally likely. Now the sample histogram can be formally defined as follows.

Definition 2.4.4. *Let X be a discrete random variable of k possible outcomes with probabilities (p_1, p_2, ..., p_k). Let $\{w_1, w_2, \ldots, w_k\}$ be the set of corresponding outcome counts when a sample of size m is taken from X, so that $\sum_{i=1}^{k} w_i = m$. Then, the sample histogram of X is defined as the k-tuple $(w_1, w_2, \ldots, w_k)$.*

A sample histogram may or may not have all k possible labels represented. In a sample, lack of representation of an outcome − producing some empty buckets or bins

in the histogram − is generally unanticipated. In such case the calculation of MSU may be biased; we give a specific name to this situation.

Definition 2.4.5. *Given a sample histogram* $(w_1, w_2, \ldots, w_k)$ *from a multinomial distribution, the sample is said to be extreme if* $w_i = 0$ *for at least one* i*.*

Clearly, if sample size m is smaller than the number of buckets k, then we have an extreme sample. Intuitively, as m increases, the probability of getting an extreme sample decreases.

In the interest of guarding ourselves against extreme samples, it is convenient to derive a way to minimize the probability of getting an extreme sample by finding a suitable sample size m.

Lemma 2.4.1. *Let* S_m *be the space of all sample histograms built from samples of size* m *from a multinomial distribution with* k *possible outcomes, and let* Z *be the count of 0s in any one of these histograms. Then the values of* Z *partition* S_m *into* k *disjoint subsets.*

Proof. Let Z_i be any value of Z. This value groups the set of sample histograms having Z_i buckets with a count of 0. A sample histogram having this count cannot belong to any other subset of the partition, because it would then have a different number of buckets with 0 count. Hence, the k subsets are disjoint. □

The random variable Z has k possible values 0, 1, ..., $k-1$. The case Z=0 corresponds to non-extreme or representative samples. The value $Z = k$, or all buckets at 0 count, never occurs because the sum of observed frequencies in the multinomial must equal the sample size m.

The lemma allows us to separately compute the probability of each partition of S_m: now these probabilities are additive because they refer to disjoint subsets.

2.4.2 Bias caused by an extreme sample

Note that when sampling m times from a categorical variable X having k labels, the actual frequency counts f_i are used to estimate the probabilities p_i. Using the "hat" notation to represent an estimate, $\hat{p}_i = f_i/m$ for all i, where

$$\sum_{i=1}^{k} f_i = m \tag{2.19}$$

What happens when one of the f_i is zero, that is, an extreme sample is obtained Equality (2.19) still holds, but one or more of the non-zero p_i are overestimated because

this particular sample, by pure chance, has loaded some buckets more heavily at the expense of the empty bucket. This often occurs in practice when the sample size is too small, or when there is a group of non-respondent individuals all belonging to the missing category, the analyst being unaware of this.

Lemma 2.4.2. *Let X be a categorical random variable with entropy $H(X)$ as defined in equation (2.1), and let S be a sample of m observations of X. Suppose label L of the variable is missing in the sample, that is, only $k-1$ distinct labels of X are found in S. Then, the estimated entropy $\hat{H}(X)$ computed from the sample using the above estimation procedure, approximates the true entropy $H(X)$ with a bias that is a function of the missing label probability p_L.*

Proof. When a label is missing in a sample of size m, some or all of the other labels appear more frequently. That is, on average each bucket will attract extra instances, its observed frequency becoming greater than the true mp_i by an amount proportional to p_i itself. Thus, assume that each non-empty bucket receives an increase in frequency d_i where $\sum d_i = p_L$; the estimated probability is then $\hat{p}_i = p_i + d_i$ for each $i \neq L$. Then, taking expectation over the density of X,

$$\begin{aligned}
E[\hat{H}(X)] &= -\sum_{i\neq L}(p_i+d_i)\log(p_i+d_i) \\
&= -\sum_{i\neq L}(p_i+d_i)\log[p_i(1+\frac{d_i}{p_i})] \\
&= -\sum_{i\neq L}(p_i+d_i)[\log(p_i)+\log(1+\frac{d_i}{p_i})] \\
&= -\left[\sum_{i\neq L}p_i\log(p_i)+\sum_{i\neq L}d_i\log(p_i)+\sum_{i\neq L}(p_i+d_i)\log(1+\frac{d_i}{p_i})\right] \\
&= -\sum_{i=1}^{n}p_i\log(p_i)-\left[\sum_{i\neq L}d_i\log(p_i)+\sum_{i\neq L}(p_i+d_i)\log(1+\frac{d_i}{p_i})-p_L\log(p_L)\right] \\
&= H(X)+\left[p_L\log(p_L)-\sum_{i\neq L}d_i\log(p_i)-\sum_{i\neq L}(p_i+d_i)\log(1+\frac{d_i}{p_i})\right]
\end{aligned} \tag{2.20}$$

Thus the bias is the expression in brackets, where the term $p_L\log(p_L)$ contains the probability of the missing label. Note the d_i's are also related to the same probability, since they add up to p_L. This completes the proof. □

The following example explores how bias behaves when a bucket is missing.

Example. Suppose random variable X has four possible values with probabilities $\{0.12, 0.18, 0.40, 0.30\}$ respectively. Also, suppose that category L was not observed or represented on the entire sample (where $L \in \{1, 2, 3, 4\}$). Let p_L be the probability of the missing bucket. Following the reasoning of the Lemma, p_L can be distributed among the remaining cells, increasing their frequency by an amount d_i. This amount can be made proportional to the other buckets' respective probabilities; thus if for example the category whose probability equals 0.18 is missing, apportion 0.18 to increase each of the other three probabilities:

$$
\begin{aligned}
d_1 &= (0.18)\, p_1/(p_1 + p_3 + p_4) = (0.18)\, p_1/(1 - p_2) \\
d_3 &= (0.18)\, p_3/(1 - p_2) \\
d_4 &= (0.18)\, p_4/(1 - p_2)
\end{aligned}
$$

That is, $d_i = p_L\, p_i/(1 - p_L)$ is the appropriate assignment for bucket i. With this allocation pattern, the expected bias is calculated by constructing Table 2.2. In each table section, the missing bucket probability appears in **bold** face.

In Table 2.2, the expected value of bias obtained by employing equation (2.20) is presented as a function of the missing bucket probability. To test how bias would behave under different probability densities, this table is calculated again assuming a very unbalanced density $\{0.03, 0.04, 0.05, 0.88\}$ and then assuming a nearly uniform density $\{0.22, 0.24, 0.26, 0.28\}$ – for brevity, these calculations are not shown. The three graphs are displayed in Figure 2.1.

Summing up, an extreme sample that misses a bucket with large probability p_L tends to cause a bias of large absolute value when estimating $H(X)$. Correspondingly, if the probability of a missing category is small then the resulting bias will tend to be small.

Also notice that the size of the bias can be important with respect to the true entropy (1.8622 for the first example, applying equation (2.1)). In the rest of the subsection it is explored how this affects estimation of the MSU.

Definition 2.4.6. *Consider the ratio $R = X_1/X_2$ of two random variables X_1 and X_2. Let $\hat{X}_1$ and $\hat{X}_2$ be estimates obtained from sample data. The* ***natural estimate*** *of R is defined as the ratio of estimates, $\hat{R} := \hat{X}_1/\hat{X}_2$.*

It is known that the expectation of a natural estimate, that is, the ratio of expectations, is only a first approximation to the true expectation of the ratio (Curtiss, 1941); but the natural estimate is often used in practice for simplicity. Let us now extend the

Table 2.2: Expected bias in $H(X)$ as a function of p_L

i	p	L	d	$d\log(p)$	$(p+d)\log(1+\frac{d}{p})$	$p_L\log(p_L)$	Expected Bias
1	**0.12**	1	0	0	0	-0.36707	
2	0.18	1	0.02455	-0.06072	0.03772321		
3	0.4	1	0.05455	-0.07211	0.08382935		
4	0.3	1	0.04091	-0.07106	0.06287201		
							-0.347605196
1	0.12	2	0.02634	-0.08058	0.04189817		
2	**0.18**	2	0	0	0	-0.44531	
3	0.4	2	0.0878	-0.11607	0.13966058		
4	0.3	2	0.06585	-0.11439	0.10474543		
							-0.420578789
1	0.12	3	0.08	-0.24471	0.14739312		
2	0.18	3	0.12	-0.29687	0.22108968		
3	**0.4**	3	0	0	0	-0.52877	
4	0.3	3	0.2	-0.34739	0.3684828		
							-0.376760476
1	0.12	4	0.05143	-0.15731	0.08821254		
2	0.18	4	0.07714	-0.19085	0.13231882		
3	0.4	4	0.17143	-0.22662	0.29404181		
4	**0.3**	4	0	0	0	-0.52109	
							-0.460885953

above Lemma's result to this natural estimate, employed as a first approximation to $\mathrm{MSU}(X_{1:n})$ in the presence of a sample.

Lemma 2.4.3. *Let S be a sample of m observations from the categorical random variables $X_1, \ldots, X_n$. Suppose one of the labels is missing in the sample for a variable X_j, where $j \in [1, n]$, causing the estimation of $H(X_j)$ to be biased. Then the natural estimate $\hat{R}$ of MSU$(X_{1:n})$ obtained by direct substitution of each entropy by its estimate is also biased and*

$$E(\hat{R}) = \frac{n}{n-1}\left[1 - \frac{E(\hat{H}(X_{1:n}))}{E(\sum_{i=1}^{n}\hat{H}(X_i))}\right] \leq \frac{n}{n-1}\left[1 - \frac{H(X_{1:n})}{\sum_{i=1}^{n}H(X_i)}\right]. \tag{2.21}$$

Proof. Let B be the bias in the estimation of $H(X_j)$, so that $E[\hat{H}(X_j)] = H(X_j) + B$. Recall from the previous lemma that $B > 0$. Working with the left hand side numerator

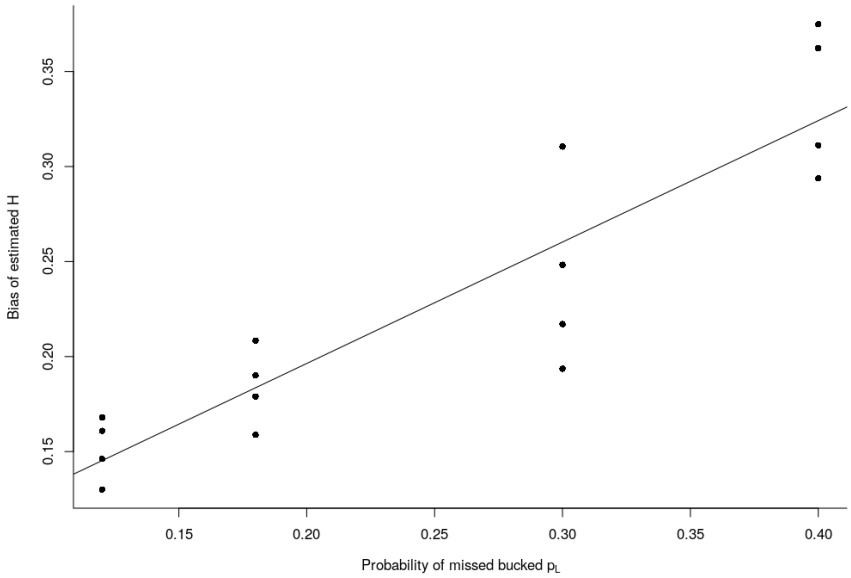


Figure 2.1: Bias in estimated H as a function of missed category probability.

in (2.21) and using the chain rule in (2.9), one has

$$\begin{aligned} E[\hat{H}(X_{1:n})] &= E[\hat{H}(X_1) + \ldots + \hat{H}(X_j) + \sum_{i=j+1}^{n} \hat{H}(X_i|X_1, \ldots, X_{i-1})] \\ &= E[\hat{H}(X_1)] + \ldots + E[\hat{H}(X_j)] + E\left[\sum_{i=j+1}^{n} \hat{H}(X_i|X_1, \ldots, X_{i-1})\right] \\ &= H(X_{1:n}) + B \end{aligned} \tag{2.22}$$

Similarly, working with the corresponding denominator in (2.21),

$$E\left[\sum_{i=1}^{n} \hat{H}(X_i)\right] = \sum_{i=1}^{n} E[\hat{H}(X_i)] = \sum_{i=1}^{n} H(X_i) + B. \tag{2.23}$$

Substituting both expectation results on the left hand side of (2.21),

$$\frac{n}{n-1}\left[1 - \frac{E(\hat{H}(X_{1:n}))}{E(\sum_{i=1}^{n} \hat{H}(X_i))}\right] = \frac{n}{n-1}\left[1 - \frac{H(X_{1:n}) + B}{\sum_{i=1}^{n} H(X_i) + B}\right] \tag{2.24}$$

Now given two positive numbers a, b such that $a \leq b$, then for any positive number c it is easy to prove that $a/b \leq (a+c)/(b+c)$ with equality holding only when $a = b$.

Recalling Lemma 2.3.1, it is known that $H(X_{1:n}) \leq \sum_{i=1}^{n} H(X_i)$. Thus

$$\begin{aligned} \frac{H(X_{1:n})}{\sum_{i=1}^{n} H(X_i)} &\leq \frac{H(X_{1:n}) + B}{\sum_{i=1}^{n} H(X_i) + B} \\ \frac{n}{n-1}\left[1 - \frac{H(X_{1:n})}{\sum_{i=1}^{n} H(X_i)}\right] &\geq \frac{n}{n-1}\left[1 - \frac{H(X_{1:n}) + B}{\sum_{i=1}^{n} H(X_i) + B}\right] \\ MSU(X_{1:n}) &\geq E(\hat{R}) \end{aligned} \tag{2.25}$$

and the proof is complete. □

2.4.3 Sample size m for representativeness

As seen, an extreme sample can bias the natural estimate of the MSU; therefore, a totally representative sample should be sought for whenever possible. It is not possible to completely avoid an extreme sample, but it is worth trying to reduce its probability to an acceptable level. For a multinomial density, if outcome count W_i in the sample histogram is observed over m trials, then the expectation and the variance are $E(W_i) = mp_i$ and $V(W_i) = mp_i(1 - p_i)$ respectively (Thompson, 1987).

Since the density type is multinomial, it is possible to establish a $1 - \alpha$ (say 95%) confidence interval that the observed frequency w_i will be close to its expected value mp_i and away from zero (a zero frequency implies an extreme sample). This can be done by imposing the condition that 0 should be outside of the confidence interval.

A confidence interval for W_i at the $1 - \alpha$ level can be constructed by noting that the observed count w_i estimates the "true" count $E(W_i)$. Hence, using the normal approximation, the following inequalities should hold with probability $1 - \alpha$:

$$-z_{\alpha/2}\sqrt{V(W_i)} \leq w_i - E(W_i) \leq z_{\alpha/2}\sqrt{V(W_i)} \tag{2.26}$$

for a two-tail interval, and

$$-z_{\alpha}\sqrt{V(W_i)} \leq w_i - E(W_i) \tag{2.27}$$

for a left-sided one-tail interval. Because we are trying to get 0 into the left tail with probability α, it makes sense to use the one-sided confidence interval. By transposing $E(W_i)$ and then adding the greater than zero condition,

$$\begin{aligned} 0 &< E(W_i) - z_{\alpha}\sqrt{V(W_i)} \leq w_i, \\ 0 &< mp_i - z_{\alpha}\sqrt{mp_i(1 - p_i)} \end{aligned} \tag{2.28}$$

Solving for m one arrives at

$$m > z_\alpha^2 \, \frac{1 - p_i}{p_i} \quad \forall \; i. \tag{2.29}$$

For a 95% one-sided interval, z_α is 1.645, the standard normal distribution value leaving a tail with area equal to 0.05.

Note that having a multinomial with all outcomes equally likely, in (2.29) the recommended sample size m will stay constant for all buckets. If, however, there are unequal p_i values, m will attain its largest value at the smallest p_i (which can be called the *least likely bucket*), and this becomes the recommended sample size.

Table 2.3: Sample size m for various multivariate cardinalities assuring total representativeness at $1 - \alpha$ level.

Univariate Cardinalities	Multivariate Cardinality k	p_i in each bucket	m for $\alpha = 0.05$	m for $\alpha = 0.01$
2, 2, 2	8	0.125	19	35
2, 2, 3	12	0.0833	30	55
2, 2, 4	16	0.0625	41	75
2, 2, 5	20	0.05	52	95
2, 3, 3	18	0.0555	47	85
2, 2, 2, 3	24	0.0416	63	115
2, 2, 2, 4	32	0.03125	84	154

Referring back to the X_1, X_2 and X_3 example at the beginning of section 2.4.1, with univariate cardinalities of 2 in each RV, there are 8 possible equiprobable outcomes so that $p_i = 1/8$ for all i. Applying (2.29) it is obtained

$$m > (1.645)^2 \, \frac{1 - \frac{1}{8}}{\frac{1}{8}} = 18.94 \approx 19 \tag{2.30}$$

Thus, a sample of size 19 gives 95% confidence that this sample histogram will not bring an extreme sample, that is, the sample will be totally representative. Table 2.3 displays a few examples with various univariate cardinalities at source features, and recommended m values for one-sided 95% and 99% assurance of total representativeness.

2.4.4 A simplified empirical expression

The example shown in equation (2.30), with bucket probability $p_i = 1/8$, reminds us that p_i is actually the inverse of the multivariate cardinality of the group of features

– the product of all cardinalities. Denoting this product by π, and substituting in equation (2.29) one has

$$\begin{aligned} m &> z_\alpha^2 \, \frac{1-p_i}{p_i} \quad \forall i \\ &= z_\alpha^2 \, \frac{1-(1/\pi)}{1/\pi} = z_\alpha^2 \, (\pi - 1) \end{aligned} \tag{2.31}$$

Setting the probability of extreme sample to a very small value, say $\alpha = 0.001$, induces a conservative (or larger) sample size. Under that condition, normal density tables give $z_\alpha^2 = (3.09)^2$ and thus

$$m > (3.09)^2(\pi - 1) \approx 10\pi \tag{2.32}$$

This is a simplified way to obtain the sample size. It agrees with a first successful attempt at finding sample size during early runs of experiments, in a way similar to Figure 2.3 in the Computational Results below. Let us name equation (2.32) as "the empirical expression 10π". Being on the conservative side allows for robust results even under noisy data, as it will be visualized in that.

2.4.5 Remarks

A multinomial density with unequal p_i values will arise when one or more categories of an underlying feature X_j have uneven probabilities. In such an event, the correct approach is to calculate the sample size for the least likely bucket, that is, the lowest p_i value in the multinomial density under consideration. The reason being that the frequency observed in this bucket tends to be low, requiring a larger m to pave the way for representativeness.

In Table 2.3 relatively low univariate feature cardinalities are shown, since this is usually the situation when dealing with categorical variables − perhaps because the human mind is comfortable handling only a few nominal categories. Often, instead of using many categories man tends to develop numeric measures.

The initial assumption of independence of the original or source features is just a special case. In the case of independent features, knowing the values of one or more features does not make it any easier to predict the values of the rest of the features; this is a situation of maximum entropy. Partial dependencies, however, imply that *some* extra prediction capability is given; at an intuitive level this means that smaller sample sizes would suffice. This matter is work to be pursued in future.

In this section it was shown that the multivariate cardinality of a set of features

induces a multinomial probability density for the combined values of those features when taking a sample of size m. From here we have been able to determine the sample size m required to assure total representativeness with a desired probability level.

2.5 Data

A variety of datasets are employed in this work, depending on the aim of each experiment. The effectiveness of MSU under different scenarios and its bias are first analyzed and evaluated using synthetic datasets, which provide the benefit of a controlled environment for analyzing strengths and limitations. These datasets also lead to the discovery of the empirical expression 10π given as equation (2.32), and are used to analyze the goodness of that expression. Then, the application of MSU to the feature selection problem is evaluated on benchmark data widely used in the field.

2.5.1 Data for MSU effectiveness analysis

Synthetic datasets are used to study the performance of MSU on a variety of scenarios. The dataset is generated following the guidelines presented in (Kononenko, 1995).

In particular, the data is generated by considering, as classification rule, the *XOR* function and the target concept introduced in (Kononenko, 1995) which will be referred to as Kononenko's method (KM) from now on.

When the *XOR* function over 2 features equals the value of a third feature (for instance the class), this can be used to test for detection of a multivariate functional dependency. This target concept is interesting because either feature has no separation power by itself; however, the two features are informative when considered together. Therefore, any algorithm that does not consider multivariate functional dependencies between features can fail on the task of selecting both features.

XOR is extensible, whereby the class may be a function of k non-redundant attributes. The class cannot be expressed as a function of a subset of those k attributes without losing its essential behavior. For our experiments, we only employ the 2-attribute version of XOR for the sake of simplicity.

KM also allows to generate equally informative features despite each having different numbers of values. This is done by joining the values of the features into two subsets: $\{1, \ldots, (V \; div \; 2)\}$ and $\{(V \; div \; 2 \; + \; 1), \ldots, V\}$. The probability that a value belongs to a subset depends on the class, while the selection of a particular value of a subset is random from the uniform distribution. The probability that the value of a feature belongs to a subset is defined as:

$$P(j \in \{1, \ldots, (\lfloor \frac{V}{2} \rfloor)\} \mid i) := \begin{cases} \frac{1}{i + kC} & \text{if } i \bmod 2 = 0 \\ 1 - \frac{1}{i + kC} & \text{if } i \bmod 2 \neq 0 \end{cases}$$

where C is the number of class labels, i is an integer indexing the possible class values $\{c_1, \ldots, c_i\}$, j is the value of the feature, and k determines how informative the feature is. A higher value of k indicates a stronger level of association between the feature and the class, making the feature more informative. However, as reported in (Kononenko, 1995), the bias of MI is not sensitive to the value of k and, therefore, all experiments in this work use $k = 1$.

2.5.2 Data for benchmarking MSU at feature selection

In order to assess MSU in the feature selection problem, four synthetic datasets widely used in feature selection literature are employed. The first dataset, Corral (John et al., 1994), contains six Boolean features ($A0$, $A1$, $B0$, $B1$, I, R) and a Boolean class $\mathcal{Y}$ defined by $\mathcal{Y} = (A0 \wedge A1) \vee (B0 \wedge B1)$. Features $A0$, $A1$, $B0$ and $B1$ are independent to each other, feature I is uniformly random, and feature R matches the class label 75% of the time. Therefore, the optimal subset includes $A0$, $A1$, $B0$ and $B1$.

The next three datasets are referred to as the Monk's problem (Thrun et al., 1992), described by six nominal features:

- Head-shape (a_1) $\in$ round (1), square (2), octagon (3)
- Body-shape (a_2) $\in$ round (1), square (2), octagon (3)
- Is-smiling (a_3) $\in$ yes (1), no (2)
- Holding (a_4) $\in$ sword (1), balloon (2), flag (3)
- Jacket-colour (a_5) $\in$ red (1), yellow, green (2), blue (3)
- Has-tie (a_6) $\in$ yes (1), no (2)

The concepts to learn are the following:

- Monk-1: (head-shape = body-shape) or (jacket-colour = red). This concept is difficult to learn due to the interaction between the first two features.
- Monk-2: Exactly two of the features have their first value. This is a hard problem because of the pairwise feature interactions and the fact that only one value of each feature is useful. Note that all six features are relevant.

- Monk-3: (jacket-colour = green and holding = sword) or
  (jacket-colour $\neq$ blue and body-shape $\neq$ octagon). This dataset has 5% class noise (with label reversed).

In addition, popular real-world datasets are selected. A summary of these datasets including the significance level α associated to the respective sample representativeness is shown in Table 2.4. The first column refers to the dataset name. Next column indicates the sample size m followed by the number of features n. Then, the multivariate cardinality of the data is given. The last column presents the implied α value when solving in (2.29) for z_α, under the given m and the given multivariate cardinality. Many α values are relatively high, due to small sample sizes compared to their respective multivariate cardinalities. This implies that their datasets face higher probabilities of extreme samples when they are run.

Table 2.4: Summary of the real-world data assuring total representativeness at $1-\alpha$ level.

Dataset	m	n	MC	α
Haberman	306	3	24	$1.3e-4$
Balance Scale	625	4	81	0.003
Iris	150	4	144	0.153
Nursery	12960	8	12960	0.159
Diabetes	768	8	1536	0.240
Heart StatLog	270	13	1024	0.304
Glass	214	9	2304	0.380
Heart H	294	13	18432	0.450
Breast Cancer	286	9	299376	0.488
Wine	178	13	746496	0.494
Sonar	208	60	2097152	0.488
Creadit A	690	15	8294400	0.496
Zoo	101	17	13107200	0.499
Lymph	148	18	28311552	0.499

2.6 Computational results

The set of experiments carried out in this chapter have the following objectives. The first experiment demonstrates MSU's ability to measure multivariable interactions. The second and third sets of experiments aim at determining what factors cause bias in MSU and how bias can be controlled. The fourth set explores MSU as an aid for

feature selection. In particular, in order to achieve these objectives, experimentation is performed as follows:

1. *Analyze the ability of MSU to detect interactions.* To this aim, experiments are performed in two scenarios:
   - Scenario 1. Using the KM, compare the values of MSU for two features and the class with values of SU for each feature with respect to the class.
   - Scenario 2. Using a XOR rule, study whether MSU is able to capture the interaction of features.
2. *Analyze the bias of MSU* due to the following factors:
   - Cardinality. Examine the behavior of MSU when varying the cardinality of the features. In this experiment the subset consists of two features − one irrelevant and the other individually informative − and the target rule is KM.
   - Dimensionality. This experiment considers the XOR scenario and computes the MSU when adding irrelevant and individually informative features. The robustness of MSU is also analyzed by adding the following noise levels: 5, 10, 15, 20 and 25 percent.
   - Sample size. The previous experiment is repeated with a small sample size.
3. *Analyze MSU behavior with a calculated sample size*: Expression 10π in equation (2.32) is used to calculate the minimum sample size to avoid extreme samples with about 0.999 confidence. In this scheme the following studies are carried out:
   - In the XOR scenario, in which irrelevant as well as individually informative features are added, report:
     - The *population* or real MSU values against those calculated with the sample size given by 10π.
     - The robustness of MSU in the presence of noise at the 5, 10, 15, 20 and 25 percent levels.
   - Compare the population MSU value with the MSU for various combinations of feature types.
4. *Assess MSU applied to feature selection*:
   - With synthetic data, the scenario allows to assess MSU in a controlled environment.

- With real-world data, MSU is evaluated as a new alternative in feature selection.

In the next four sections, each experiment is presented in detail.

2.6.1 Analyze the ability of MSU to detect interactions

In order to assess the capability of MSU to detect interactions among features, two scenarios are considered. In the first one, the results of SU on informative and irrelevant features are compared with the results obtained by MSU on the same variables. In the second scenario, attention is focused on the capabilities of MSU to capture the interaction between features.

2.6.1.1 Scenario #1

The aim of this experiment is to compare the results of MSU on two features, one informative and one irrelevant, with the result obtained by SU on the same features with respect to the class. In all cases the feature cardinalities are varied from 2 to 50 and class cardinality is fixed at 2. This experiment was performed using a sample size of 10^5 in order to minimize any effect associated to the sample size.

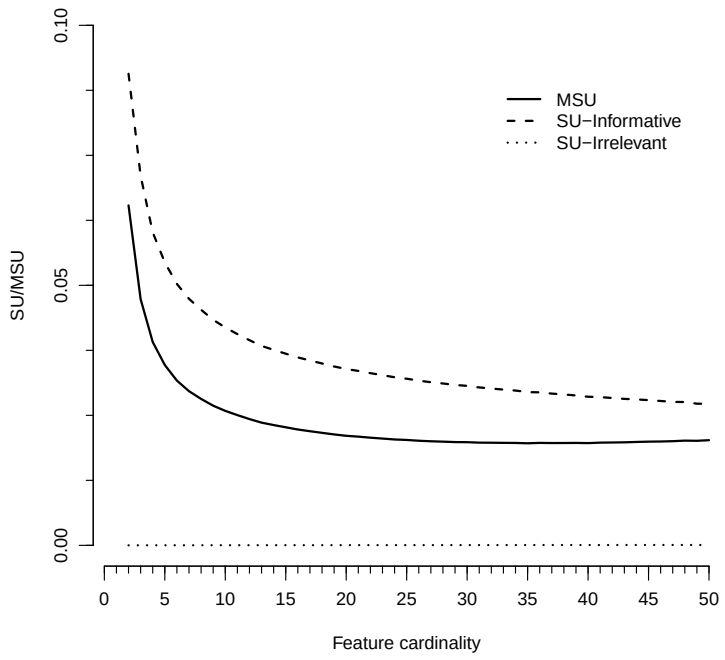


Figure 2.2: SU and MSU for different values of feature cardinality and dichotomous class.

Figure 2.2 presents the comparison of SU and MSU values. SU has a value close to 0 for the irrelevant feature regardless of the cardinality. For the informative feature, SU decreases when cardinality of the feature increases. In this case, the SU slope flattens as cardinality increases. The MSU curve presents a similar trend, with a lower level because of the irrelevant feature.

2.6.1.2 Scenario #2

The capability of MSU to capture the interaction between features is studied in this scenario. The dataset used in this experiment is composed of two binary features and the target rule for classification is the XOR function. Also, the effect of the sample size on MSU is analyzed by varying the number of instances from 8 to 150. The robustness in the presence of noise is analyzed by adding various levels of white noise to the features. In particular, the following percentage levels of noise were considered: 0, 5, 10, 15, 20 and 25. To gain insight on these issues, two experiments are performed:

- In the first experiment, it is analyzed the ability of MSU for capturing the interaction of two features that are individually non-informative, but are collectively informative as it is the case of the XOR rule.
- The second experiment is similar, adding levels of white noise to analyze the robustness of MSU in the presence of noise.

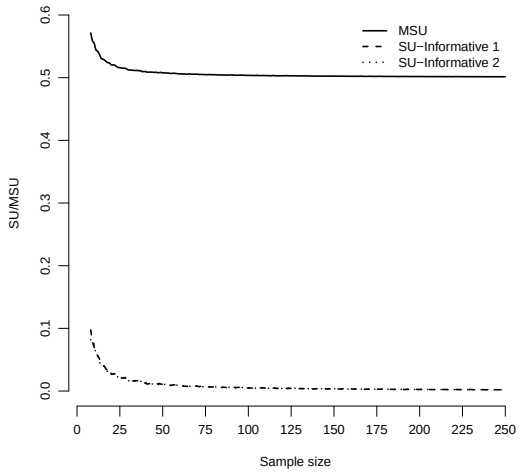
The results of the first experiment in this scenario are shown in Figure 2.3a. SU values of individually non-informative feature are low and converge to 0 as the sample size increases. In contrast, MSU always assumes higher values. This is due to the fact that MSU is able to capture the interaction of both features, converging to the "true" value of 0.5 for large sample sizes. Figure 2.3b presents the values of MSU for different levels of noise. As it can be expected, the higher the noise levels, the lower the MSU values.

2.6.2 Analyze the bias of MSU

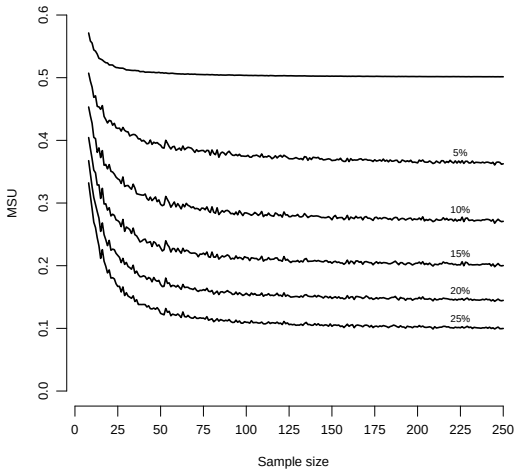
In this section, experiments are presented so as to study the effect that several factors (cardinality, dimensionality and sample size) have on the MSU.

2.6.2.1 Analysis of cardinality bias

A first experiment analyzes how the cardinality influences MSU. To this end, a synthetic dataset of two features (one informative and the other irrelevant) is used, comparing



(a) SU and MSU values of collectively informative features that follow the XOR function as target rule.



(b) Robustness of MSU to capture the XOR function at different noise levels.

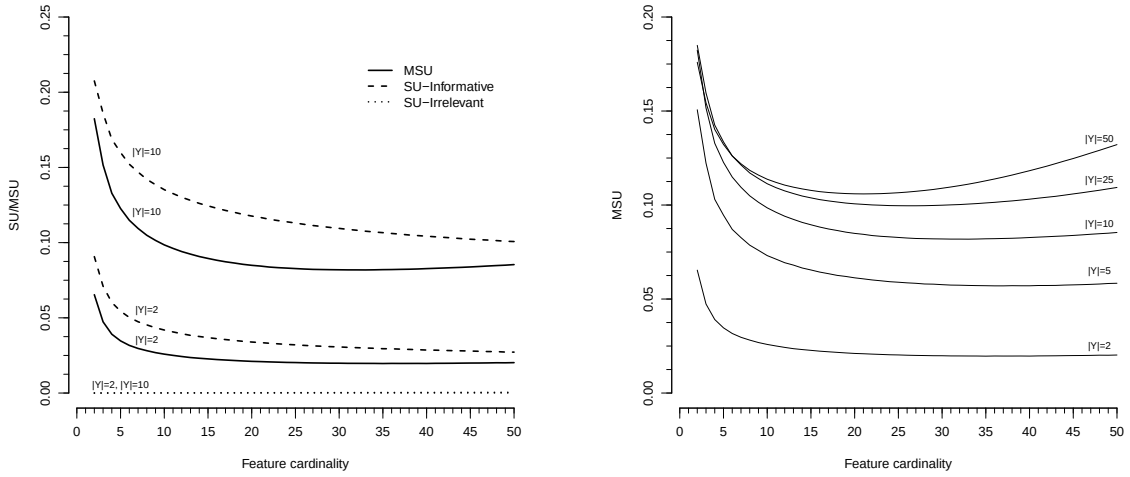
Figure 2.3: Study of the capability of MSU to capture interaction of features, and its robustness in presence of noise.

the results obtained by MSU with those achieved by SU on each feature individually. In order to examine the effect of the cardinality on the measures, the number of values of both features and the class varied from 2 to 50. In all cases the sample size was fixed at 10^5 to minimize any effect associated to the sample size.

Figure 2.4 shows the results of these experiments. In particular, Figure 2.4a compares SU and MSU while Figure 2.4b presents the values of MSU obtained under various class cardinalities.

In Figure 2.4a, we can observe the values of SU and MSU for two cardinalities of the class ($|Y| = 2$ and $|Y| = 10$, augmenting on Figure 2.2). In both cases, the irrelevant feature is characterized by SU values close to 0. As far as the informative feature is concerned, the behavior is similar for both class cardinalities: given a class cardinality value, SU decreases when the number of values of the informative feature increases. Also, higher class cardinalities produce higher values of SU. The behavior of MSU is similar to that of SU for the informative case; and again, MSU values are slightly lower than the SU because the irrelevant feature "pulls" them down.

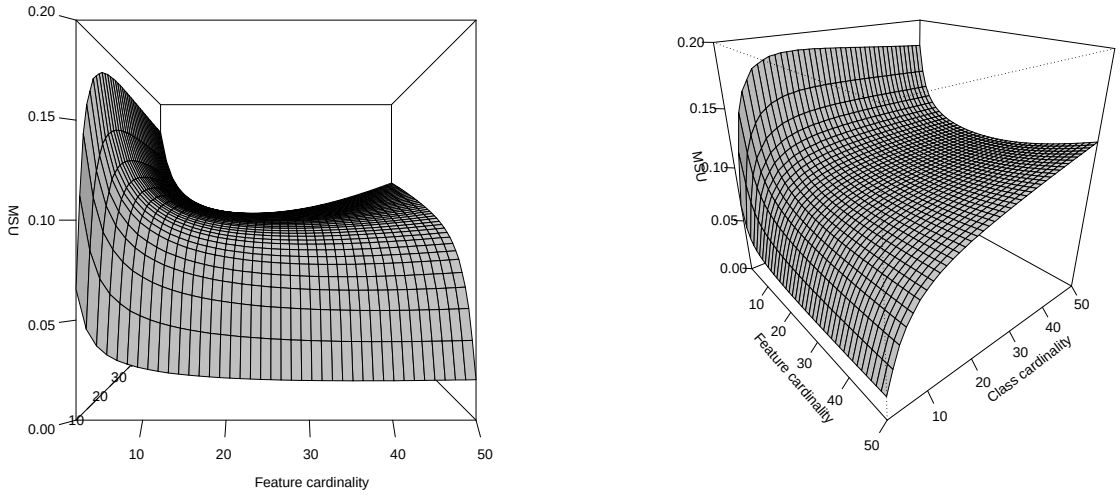
Figure 2.4b plots an MSU curve for each class cardinality, with increasing informative feature cardinality. Higher values of either cardinality would require larger sample size to control the bias. For our sample of 10^5, a cardinality of 10 shows a small bias but this effect increases for higher cardinalities.



(a) SU and MSU for the class cardinalities of 2 and 10.

(b) MSU for different class cardinalities with a fixed sample size of 10^5.

Figure 2.4: Effect of varying class cardinality on SU and MSU. Sample size is fixed at 10^5.



(a) **(b)**

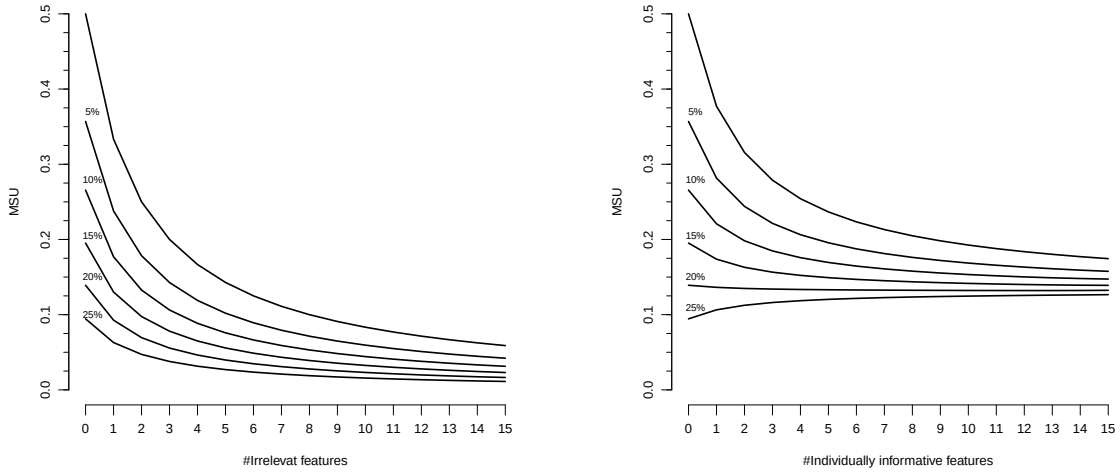
Figure 2.5: Effect of varying class cardinality on MSU. Sample size is fixed at 10^5.

In order to gain a better perspective of how MSU values change when increasing feature and class cardinalities, two perspectives are shown of the corresponding three-dimensional surface in Figure 2.5. At a fixed sample size of 10^5, increasing class cardinality sharply increases MSU values specially at the beginning; whereas increas-

ing feature cardinality causes MSU to quickly decrease and then become stable after reaching a cardinality of about 20.

2.6.2.2 Analysis of dimensionality bias

In this section, the bias of MSU associated with dimensionality is analyzed. In order to do so, the dataset generated with the XOR function is employed to examine how the value of MSU changes when adding irrelevant and individually informative features. In order to avoid any bias due to any other factor, features with a cardinality of 2 and a sample size of 10^8 are considered. The robustness of MSU against noise is also tested by setting different levels of noise, ranging from 0% to 25%, as in previous sections.



(a) MSU values when adding irrelevant features.

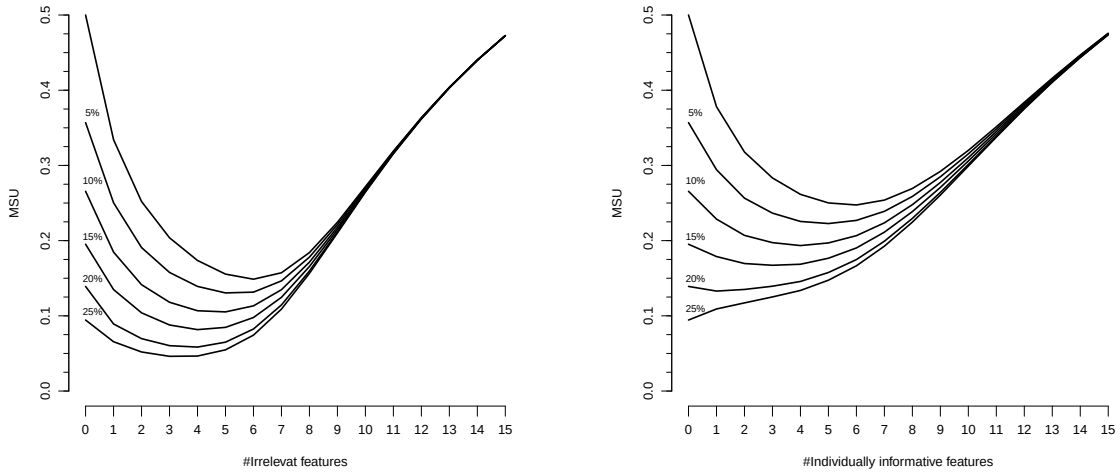
(b) MSU values when adding individually informative features.

Figure 2.6: MSU bias due to dimensionality. In all cases the sample size is fixed to 10^8.

Figure 2.6 presents the values of MSU when adding irrelevant features (Figure 2.6a) and individually informative features (Figure 2.6b). In the first case the addition of features results in an asymptotic decrease in MSU values towards 0. In the second case the addition produces a slow convergence towards a "real" level of correlation induced at KM's generation of informative features. Furthermore this behavior is similar when increasing the noise – higher noise levels imply lower MSU values.

2.6.2.3 Analysis of sample size bias

The bias associated to size of the sample is explored in this section. It is shown that a small sample size implies increasing MSU values as the number of features increases.



(a) Effect of sample size when adding irrelevant features.

(b) Effect of sample size when adding individually informative features.

Figure 2.7: Effect of sample size on MSU. The sample size is fixed to 1000.

Again, considering the XOR scenario, the previous experiment of adding irrelevant features is repeated, with the difference that a small sample of 1000 is used. Figure 2.7 corresponds to the addition of irrelevant and individually informative features (see 2.7a and 2.7b) respectively. As one can see, MSU becomes strongly biased upwards when dimensionality increases.

2.6.3 Analyze MSU behavior with a calculated sample size

In this section, the behavior of the measure is analyzed by comparing the values of MSU from a very large dataset with those from a sample fixed to the size calculated by expression 10π. Continuing to use the XOR scenario, the following experiments are performed:

- First, the values achieved by MSU when adding irrelevant and individually informative features in a population of 10^8 are compared with those obtained from a sample size calculated by the empirical expression 10π. Then, the robustness in both cases is studied by adding noise at percentual levels $\{5, 10, 15, 20\}$.
- Second, MSU values are calculated for subsets of mixed types of features for a population of 10^6 and for sample sizes calculated by 10π, and the results are compared.

2.6.3.1 MSU and addition of features

In two sets of experiments, the values achieved by MSU when adding irrelevant and individually informative features in a population of 10^8 are compared with the values from a sample whose size is calculated by the empirical expression 10π. Then, the robustness in both cases is studied by adding noise at percentual levels $\{5, 10, 15, 20\}$. Figure 2.8 shows the results of both scenarios. The comparison between the population MSU and the sample MSU when adding irrelevant features is shown in Figure 2.8a, with the values of MSU very close to population MSU in all cases. Irrelevant features with noise are shown in Figure 2.8b, where MSU behaves in a similar way but values decrease when increasing noise levels. Results with individually informative features are very similar (Figures 2.8c and 2.8d), but achieving higher MSU values than with irrelevant features.

2.6.3.2 MSU values with mixed types of features

In this experiment, MSU values are calculated for subsets of 2 to 9 features of mixed types. Values for populations of 10^6 are compared with values obtained from samples with sizes calculated by expression 10π. For each subset, the types of features were randomly chosen.

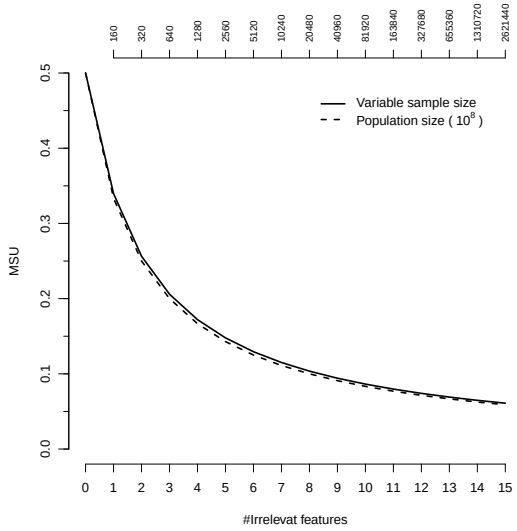
The paired histogram in Figure 2.9a shows the value of MSU on the population and on the sample, for each subset of features. The number of instances in each subset is reported in the upper side. The types of features of each subset are shown in Figure 2.9b. Overall, MSU values estimated from samples are quite close to the true population values, except in the case of the 2 non-informative attributes where the true MSU value is 0.

2.6.4 Assess MSU applied to feature selection

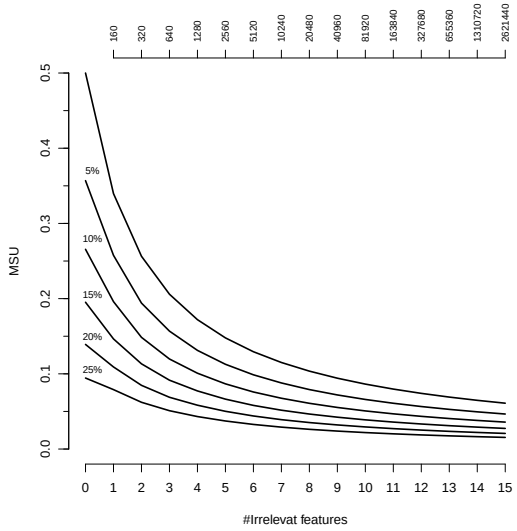
In the following, experiments aimed at assessing the ability of MSU to evaluate subsets of features are presented. Results obtained on synthetic data are presented first, and then the results achieved on real-word datasets.

2.6.4.1 Results on synthetic data

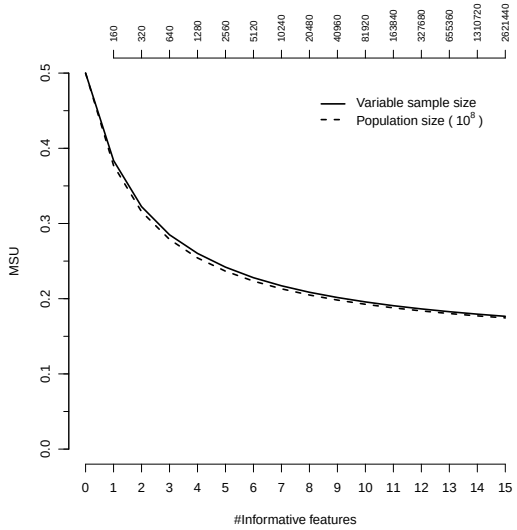
In this experiment, an exhaustive search is applied to report the subset with the highest MSU value using four synthetic datasets: Corral, Monk-1, Monk-2 and Monk-3 which were introduced in the Data chapter. Results are shown in Table 2.5, where the first column refers to the dataset name followed by the target concept $\mathcal{Y}$. The third column



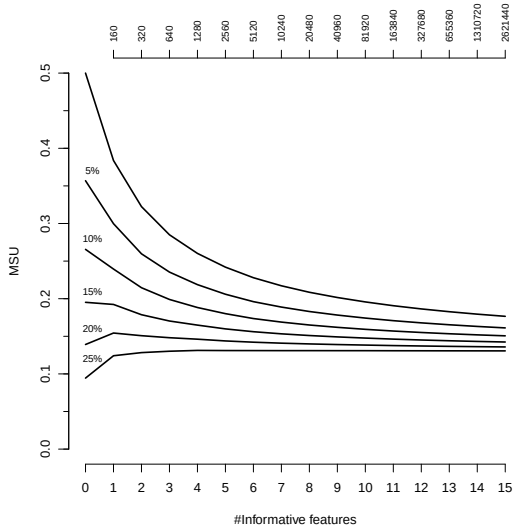
(a) Sample MSU and population MSU when adding irrelevant features.



(b) MSU values when adding irrelevant features in the presence of noise.



(c) Sample MSU and population MSU when adding individually informative features.

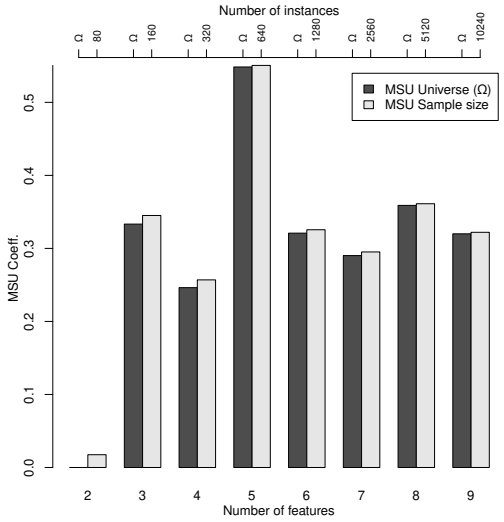


(d) MSU values when adding individually informative features in the presence of noise.

Figure 2.8: Comparison between sample MSU and population MSU, and analysis of the effect of noise in MSU. Sample sizes calculated by expression 10π are shown for each curve. Total population size is 10^8.

shows the relevant features R to be selected, finally, in the last column, the subset of features S with the highest MSU value is shown.

As can be noticed from the last two columns of the table, MSU is capable of identifying the important features in the datasets, with the exception of features a_4 for the Monk-3 dataset. A possible cause for this inaccuracy is that the dataset includes



(a) MSU of feature sets for the population and for the calculated sample size.

# attr.	Type of features
2	NI,NI
3	CI,CI,II
4	II,CI,CI,NI
5	CI,NI,CI,CI,CI
6	CI,CI,CI,II,NI,NI
7	CI,II,CI,II,NI,NI,CI
8	CI,II,CI,NI,CI,CI,NI,NI
9	CI,NI,NI,CI,II,CI,NI,CI,NI

NI: Non-Informative feature.
II: Individually-Informative feature.
CI: Collectively-Informative feature.

(b) Types of the randomly selected features conforming each set.

Figure 2.9: Comparing real MSU in a 10^6-instance universe with the MSU from a sample whose size is calculated with the proposed precision.

Table 2.5: Summary of the results of exhaustive search on synthetic datasets.

Dataset	$\mathcal{Y}$	R	S
Corral	$(A0 \wedge A1) \vee (B0 \wedge B1)$	$A0, A1, B0, B1$	$A0, A1, B0, B1$
Monk-1	$(a_1 = a_2) \vee (a_5 = 1)$	a_1, a_2, a_5	a_1, a_2, a_5
Monk-2	$\{a_i = 1 \wedge a_j = 1\}$ $i \neq j, j = 1, \ldots, 6$	$a_1 - a_6$	$a_1 - a_6$
Monk-3	$(a5 = 3 \wedge a4 = 1) \vee$ $(a5 \neq 4 \wedge a2 \neq 3)$	a_2, a_4, a_5	a_2, a_5

5% class noise as described in the Data chapter.

2.6.4.2 Results on real-world data

The results on real-world datasets are presented in this subsection. In order to assess the performance of MSU, the greedy strategy called Sequential Forward Selection (SFS) is used, taking as evaluation measures the MSU and the Correlation-Based Feature Selection (Hall, 1998) (CFS). The experiments were performed with Naïve Bayes due to its popularity and good results achieved on several real-world datasets.

Model quality on the data is assessed through a k-fold cross-validation scheme, where k is set to 10. In real-world data the true generalization error is not usually

known and, therefore, it is not possible to determine whether a given estimate is an overestimate or underestimate. However, cross-validation is suitable for model comparison purposes. The statistical t-test was also applied to support the discussion.

Table 2.6: Accuracy achieved and number of features found by SFS using MSU and CFS.

	Accuracy		Number of features	
Dataset	MSU	CFS	MSU	CFS
Haberman	73.87 ± 6.25	73.87 ± 6.25	1.00 ± 0.00	1.00 ± 0.00
Balance Scale	90.72 ± 1.98	90.72 ± 1.98	4.00 ± 0.00	4.00 ± 0.00
Iris	94.67 ± 5.26	96.00 ± 5.62	1.40 ± 0.52	2.00 ± 0.00
Nursery	70.97 ± 0.83	70.97 ± 0.83	1.00 ± 0.00	1.00 ± 0.00
Diabetes	74.22 ± 2.47	76.57 ± 2.91	1.70 ± 2.21	3.40 ± 0.52
Heart Statlog	72.22 ± 12.50	84.07 ± 6.06	2.80 ± 3.55	6.50 ± 1.18
Glass	50.91 ± 5.51	48.61 ± 5.40	7.00 ± 0.00	6.50 ± 0.71
Heart H	82.94 ± 6.57	83.31 ± 5.04	2.50 ± 0.53	3.20 ± 0.42
Breast Cancer	74.13 ± 6.64	71.69 ± 7.20	8.40 ± 1.90	4.20 ± 0.92
Wine	81.86 ± 12.46	96.67 ± 5.37	1.80 ± 0.92	8.20 ± 1.23
Sonar	67.74 ± 11.40	65.33 ± 11.13	9.90 ± 7.92	17.70 ± 0.95
Credit A	85.51 ± 4.73	85.51 ± 4.73	1.00 ± 0.00	1.00 ± 0.00
Zoo	98.00 ± 4.22	95.00 ± 7.07	9.10 ± 0.32	9.60 ± 1.43
Lymph	76.29 ± 10.77	75.52 ± 12.53	9.50 ± 8.96	8.20 ± 2.30
Mean	78.15	79.56	4.36	5.46

Table 2.6 shows the results obtained by SFS using both MSU and CFS. On average, the accuracy achieved is similar on all datasets, being slightly higher for CFS. Only in Heart Statlog and Wine dataset the classification models learned with MSU are worse than those obtained when CFS was used. However, these differences are not statistically significant according to a t-test (p-value = 0.77).

Table 2.6 also shows the number of features selected by SFS when using MSU and CFS. It can be noticed that, on average, the use of MSU yields to the selection of fewer features. In the cases of the Heart Statlog, Wine and Sonar datasets, smaller subsets of features were selected when SFS used MSU, while the opposite is true on the Breast Cancer dataset. As for the accuracy, also in this case such differences are not statistically significant, since the p-value is 0.49.

2.7 Conclusions and future work

In this chapter, we introduce the Multivariate Symmetrical Uncertainty (MSU) measure, as an extension of the Symmetrical Uncertainty (SU) to the multivariate case. In order to evaluate the proposal, several experiments on synthetic datasets are conducted. Results confirm that MSU is a reliable multivariate correlation measure for nominal variables, with promising properties, capable of detecting linear and non-linear dependencies or interactions.

Three important factors that contribute to the bias of MSU are identified, namely the dimensionality, the cardinality of features and the sample size. In addition, it is experimentally observed that the effects of high dimensionality and high cardinalities are controllable by using larger sample sizes. Based on this, a condition is derived that allows the determination of the proper number of samples to avoid bias with a desired probability $1 - \alpha$. For such purpose, the concepts of *total representativeness* and *extreme sample* are introduced. The former defines how accurately a given sample reflects the entire population while the latter specifies a situation where one or more feature categories are missing from the observed outcomes in the sample.

Since the observed robustness and properties of the proposed MSU measure have been established, an assessment on a feature selection problem for classification tasks is performed through several experiments on both synthetic and real-world datasets. Results on synthetic data reinforced previous conclusions that MSU is capable of capturing interaction of two or more features. On real-world data, results show that MSU can be used as a feature subset evaluator method capable of finding subsets of relevant features. These subsets yield comparable classification accuracy for similar numbers of features on most datasets with respect to the CFS (Correlation-Based Feature Selection) strategy.

MSU accuracy depends on samples that are totally representative. Sample sizes as presented in this chapter are based on the premise of independence among source features, an assumption made for the analysis purposes of this work. This assumption can be relaxed in future studies to allow for previously known partial dependencies, and achieving greater generality in solutions, all of which may lead to smaller required sample sizes.

Chapter 3

Feature Selection: A perspective on inter-attribute cooperation.

High-dimensional datasets depict a challenge for learning tasks in data mining and machine learning. Feature selection is an effective technique in dealing with dimensionality reduction. It is often an essential data processing step prior to applying a learning algorithm. Over the decades, filter feature selection methods have evolved from simple univariate relevance ranking algorithms to more sophisticated relevance-redundancy trade-offs and to multivariate dependencies-based approaches in recent years. This tendency to capture multivariate dependence aims at obtaining unique information about the class from the intercooperation among features. This chapter presents a comprehensive survey of the state-of-the-art work on filter feature selection methods assisted by feature intercooperation, and summarizes the contributions of different approaches found in the literature. Furthermore, current issues and challenges are introduced to identify promising future research and development.

3.1 Introduction

Large amounts of data are being generated in various fields of scientific research, including economic, financial, and marketing applications (Chanda et al., 2009). These data often have the characteristic of high dimensionality, which poses a high challenge for data analysis and knowledge discovery. Redundant and irrelevant features increase the learning difficulty of the prediction model, cause overfitting and reduce prediction performance (Yao et al., 2022). In order to use machine learning methods effectively, preprocessing of the data is essential. Feature selection has been proven effective in preprocessing high-dimensional data and in enhancing learning efficiency, from both

theoretical and practical standpoints (Blum and Langley, 1997; Liu and Motoda, 2012; Guyon and Elisseeff, 2003). Thus, to overcome problems arising from the high dimensionality of data, feature selection removes irrelevant and redundant dimensions by analyzing the entire dataset.

Depending on whether the class label is used in the feature selection process or not, the feature selection methods can be categorized into supervised and unsupervised. Unsupervised feature selection is used to explore the dataset without the labeled data. The supervised feature selection uses the labels of samples to select the feature subset. In addition, supervised feature selection methods are usually grouped into three main categories: wrapper, embedded, and filter methods (Guyon and Elisseeff, 2003; Liu et al., 2010; Liu and Zhao, 2012; Zhong et al., 2004).

Wrappers search the space of feature subsets, using the classifier accuracy as the measure of utility for a candidate subset (Kohavi and John, 1997; Wan et al., 2022). The main advantage of such an approach is that the feature selection phase benefits from the direct feedback provided by the classifier. However, there are clear disadvantages in using the wrapper approach. The computational cost is huge, while the selected features are specific for the considered classifier. Embedded methods (Guyon et al., 2008) select features by determining which features are more important in the decisions of a predictive model. The wrapper and embedded methods can be categorized as classifier-dependent. On the other hand, strategies based on the filter approach can be categorized as classifier-independent (Macedo et al., 2019).

The filter method approach evaluates the features' relevance based on the data's intrinsic properties, being independent of the learning process. In general, filters are relatively inexpensive in terms of computational efficiency; they are simple and fast, and, therefore, most of the designed methods pertain to this category (Bolón-Canedo et al., 2016). Furthermore, in real-world applications, many of the most frequently used feature selection algorithms are also filters (Liu et al., 2010).

Recently, hybrid and ensemble methods were added to the general framework of feature selection in order to take advantage of both filter (computational efficiency) and wrapper (high performance) approaches (Almugren and Alshamlan, 2019).

Over the decades, filter feature selection methods have evolved from simple univariate relevance ranking algorithms to more sophisticated relevance-redundancy trade-offs and to a multivariate dependencies-based approach in recent years. We refer to the latter as *cooperativeness* (also known as complementariness (Chen et al., 2015), synergy (Zeng et al., 2015a) and interaction (Jakulin and Bratko, 2003b)).

Cooperating features are those that individually appear to be irrelevant or weakly

relevant to the class; but taken in combination with other features, they are highly correlated to that target class.

The simplest example is probably the behavior of a XOR-patterned database of 3 attributes X_1, X_2 and a class C. Since in this case $SU(X_1, C) = SU(X_2, C) = 0$, one is tempted to conclude that both X_1 and X_2 are irrelevant with respect to C. However, $MSU(X_1, X_2, C) > 0$; hence X_1 and X_2 intercooperate to determine the value of C. As a result a first simple rule for finding intercooperations can be "find attributes X such that $SU(X, C)$ equals 0 or nearly 0, then pair each of these with C to check their relevance with respect to the class."

In particular, this relates to the fact that ignoring possible feature interdependencies results in subsets with redundancy and lack of cooperative features (Guyon and Elisseeff, 2003; Jakulin and Bratko, 2004), which in turn cannot achieve optimal classification performance in most domains of interest (Xue et al., 2015).

Finding relationships and dependencies among variables (that is, features and/or class) is usually accomplished by employing some measure. These relationships are relevance, redundancy, and cooperativeness (the latter being viewed as interaction, complementarity, or synergy). Generally, the filter methods are based on these concepts (Vergara and Estévez, 2014).

Several studies (see Section 3.4) showed that taking into account high-order dependencies among variables can improve the performance of feature selection. More recently, Wan et al. (Wan et al., 2022) proposed a feature selection strategy using a filter-wrapper approach called $R2CI$, which takes into account multiple-feature correlations. In this paper (Wan et al., 2022), the observations made on multiple dependencies are particularly interesting as they characterize complementarity and interaction, from the point of view of the two subsets of attributes (selected and non-selected) that are being generated into the search space. Undoubtedly, feature intercooperation has been drawing more attention in recent years. Thus, the literature on cooperativeness-based feature selection that considers feature dependence shows an increase despite early research on interaction information dating back to McGill (1954) (McGill, 1954) and subsequently advanced by Han (1980) (Han, 1980), Yeung (1991) (Yeung, 1991), Tsujishita (1995) (Tsujishita, 1995), Guyon and Elisseeff (2003) (Guyon and Elisseeff, 2003), Jakulin and Bratko (2004) (Jakulin and Bratko, 2004) and Kojadinovic (2005) (Kojadinovic, 2005).

So, despite this research area receiving significant attention in recent years (most of the work has been published in the last decade), the problem is still challenging, and new algorithms emerge as alternatives to the existing ones.

In this chapter, we focus on filter methods for feature selection based on feature intercooperation. We provide a comprehensive survey of the state-of-the-art work, and a discussion of the open issues and challenges for future work. For all the reviewed algorithms we provide the year of their first appearance in the scientific literature; the chronological perspective of feature cooperativeness evolution is presented in this manner.

We expect this survey to attract attention from researchers working on different feature intercooperation paradigms to investigate further effective and efficient approaches to addressing new challenges in feature selection.

This review chapter is structured as follows. In the next section, we provide the basements of feature evaluation. Filter methods foundations are introduced in Section 3.3. Then, we will review the literature on feature selection methods based on feature intercooperation in Section 3.4, followed by a general discussion about issues and future challenges in Section 3.5. Finally, we present our principal conclusions and future research lines.

3.2 Feature Evaluation

Evaluation measure is a key part of feature importance criterion, which forms the basis of feature selection methods (Liu and Motoda, 2012). The feature selection objective is to find the relevant features (individually or in cooperation) and to discard redundant and irrelevant features in order to preserve the information contained in the whole set of input variables with respect to the target class.

The traditional correlation proposed by Pearson only computes the correlation between two numeric features. The ranked correlation measures by Spearman and separately by Kendall (Croux and Dehon, 2010) do the same for two ordinal or ranked variables. But in addition to features that express quantity or order, there are also qualitative features in real life. Qualitative features are more generic in the sense that every numeric attribute can be made qualitative by employing discretization methods (Lavangnananda and Chattanachot, 2017). Ordinal features are already qualitative in their nature. In this work, we look at correlations between two or many qualitative features. At present, information theory methods are the ones that allow to compute correlations between two or more qualitative attributes, thus opening the doors to research in generalized feature selection techniques.

We use information theory measures to quantify relevance, redundancy, and cooperativeness. Here, we show these concepts and basic definitions as follows.

3.2.1 Bivariate information measures

3.2.1.1 Mutual Information

Mutual information (also called information gain (IG) (Quinlan, 1993) or two-way interaction (Jakulin and Bratko, 2004)) measures the amount of stochastic dependency between variables, hence it can be used as a bivariate measure of correlation.

Definition 1. *Consider a discrete random variable X, with possible values $\{x_1, \ldots, x_k\}$ and probability mass function $P(X)$, and suppose we draw a series of X values. The entropy H of the variable X is a measure of the uncertainty in predicting the next value of X and is given by*

$$H(X) := -\sum_i P(x_i) \log_2(P(x_i)). \tag{3.1}$$

The mutual information $I(X, Y)$ measures the reduction in uncertainty about the value of X when the value of Y is known, as expressed in the next definition.

Definition 2. *For discrete random variables X and Y, the mutual information $I(X, Y)$ is*

$$\begin{aligned} I(X;Y) &:= H(X) - H(X \mid Y) \\ &= H(Y) - H(Y \mid X) \\ &= H(X) + H(Y) - H(X, Y), \end{aligned} \tag{3.2}$$

where $H(X, Y)$ is the extension of $H(X)$ using joint probabilities $P(x_i, y_j)$ in the definition of entropy.

It can be shown that $I(X; Y) = 0$ when X and Y are statistically independent.

3.2.1.2 Symmetrical Uncertainty

Because the mutual information tends to be larger for variables with more labels, it is convenient to normalize its values using both entropies, originating the Symmetrical Uncertainty (SU) measure (Press et al., 1988) expressed as

$$SU(X, Y) := 2 \left[\frac{I(X;Y)}{H(X) + H(Y)} \right]. \tag{3.3}$$

SU restricts its values to the range [0,1].

3.2.2 Multivariate information measures

3.2.2.1 Interaction Information

Interaction information (McGill, 1954) among multiple variables can be understood as the amount of information shared or bound up in a set of n random variables, but cannot be found within any subset of those n variables. Then, the interaction information among three variables (3-way interaction information) is given by

$$\begin{aligned} I(X;Y;Z) &:= I(X;Y \mid Z) - I(X;Y) \\ &= I(X;Z \mid Y) - I(X;Z) \\ &= I(Z;Y \mid X) - I(Z;Y). \end{aligned} \tag{3.4}$$

Unlike mutual information, the interaction information can be negative, positive, or zero (Jakulin, 2005).

3.2.2.2 Multivariate Symmetrical Uncertainty

To quantify the dependency among more than two variables, the Multivariate Symmetrical Uncertainty (MSU) (Sosa-Cabrera et al., 2019) has been proposed as a generalization of the SU according to the following expression.

$$MSU(X_{1:n}) := \frac{n}{n-1}\left[1 - \frac{H(X_{1:n})}{\sum_{i=1}^{n} H(X_i)}\right], \tag{3.5}$$

where $H(X_{1:n})$ is the extension of $H(X)$ using the joint probabilities of variables $X_1, \ldots, X_n$. Like the Symmetrical Uncertainty, MSU restricts its values to the range $[0, 1]$.

3.2.3 Type of dependencies in data

Generally, for the evaluation of attributes, the feature selection process is based on the typing of dependencies between variables. According to the behavior of one or several attributes (as possible predictors of the class), dependencies can be classified into univariate/multivariate relevance and univariate/multivariate redundancy. This section aims to present a concise review of these notions. In particular, we will address a special case of multivariate relevance, which we have designated as intercooperativeness.

Let F, S, and C denote the original feature set, the selected feature subset, and the target class, respectively. Through univariate and multivariate measures based on

information theory, we offer definitions of relevance, redundancy, and intercooperativeness. Also, we characterize a variable as relevant, redundant, or intercooperative for these sets.

3.2.3.1 Relevance

In the literature, several works (Bell and Wang, 2000; Caruana and Freitag, 1994; Koller and Sahami, 1996) have made an effort to classify the features according to their contribution to the meaning of the class concept. In this context, feature relevance has arisen as a measure of the amount of relevant information that a feature may contain about the class, where the level of individual relevance is defined either in terms of mutual information as $I(f_i; C)$, or $SU(f_i, C)$ using Symmetrical Uncertainty analogously.

In this context, a feature is considered irrelevant if it contains no information about the class and is unnecessary for the predictive task.

3.2.3.2 Redundancy

Redundancy is generally defined in terms of feature correlation, and thereby, it is quantifiable with the level of dependency among two or more features. In terms of information measures, a bivariate approach for feature redundancy is defined as $I(f_i; f_j)$, while a negative value of $I(f_i; S; C)$ indicates partial or complete redundancy in multivariate approach (Wang et al., 2013; Yu and Liu, 2004; Sosa-Cabrera et al., 2019). Similarly, to measure the common portion of information received from a set of features, Multivariate Symmetrical Uncertainty can be used as $MSU(f_{1:n})$ where $f_{1:n}$ represents the feature set $f_1, \ldots, f_n$.

3.2.3.3 Intercooperativeness

Intuitively, the intercooperativeness measures the amount of information received from grouped features, instead of separate features (Sosa-Cabrera et al., 2019; Vergara and Estévez, 2014; Jakulin, 2005). This concept, in which a set of features cooperate to predict the class label, can be quantified according to a positive value of the expression $I(f_1; f_2; \ldots; f_n; C)$. Similarly, $MSU(f_1, f_2, \ldots, f_n, C)$ can be used as a measure of the cooperative association of two o more features $f_1, f_2, \ldots, f_n$ along with the C class variable.

Feature cooperativeness must be measured with the target variable, that is, compute the relevance of a feature to the class with at least another feature presented.

Therefore, three-way dependency is the minimum order for the evaluation method of intercooperativeness.

Note that to describe this same concept, the terms interaction, synergy, and complementarity are used interchangeably throughout the literature. However, in Section 3.5, we shall argue for a more precise interpretation for each case.

3.3 Filter Methods

A filter feature selection process is independent of any learning algorithm and relies on underlying attributes of data. Thereby, to evaluate the utility of features, a filter model depends on statistical criteria applied to data such as distance, dependency, information, consistency, and correlation (Ullah et al., 2017).

A filter feature selection method attempts to select the minimally sized subset of features according to a loop of subset generation (by search strategy) and its evaluation (by measure) until some stopping criterion is satisfied (Cai et al., 2018). Based on these basic steps, an abstract algorithm for feature selection that shows the behavior of any filter method in a unified form is depicted in Algorithm 1.

Algorithm 1: A generalized filter method

Input: Full feature set F, a subset from which to start the search S_0, and a stopping criterion δ.
Output: most informative feature subset S_{best}.

$S_{best} \leftarrow S_0$ // initialize S_{best}.
$\gamma_{best} \leftarrow evaluate(S_0, F, M)$ // evaluate S_0 by measure M.
repeat
    $S \leftarrow search\ strategy(F, S_{best})$ // generate next candidate subset.
    $\gamma \leftarrow evaluate(S, F, M)$ // evaluate S by measure M.
    if γ *is better than* γ_{best} **then**
        $\gamma_{best} \leftarrow \gamma$ // update γ_{best}.
        $S_{best} \leftarrow S$ // update S_{best}.
    end
until δ ***is*** *reached*
return S_{best}

3.4 Feature-Intercooperation-based filter methods

In the feature selection field, the detection and significance of higher order interactions between variables have been a matter of discussion and experimentation, especially in

recent years.

In this section, we briefly summarize the existing filter feature selection methods assisted by feature intercooperation, looking at three aspects: the estimation of high-dimensional dependencies, the search techniques, and the number of higher-order interactions.

3.4.1 Estimation of high-order interactions.

Information theoretic quantities, such as mutual information and its generalizations, have several advantages as measures of multiple variable dependence. They are inherently model-free and non-parametric, and exhibit only modest sensitivity to undersampling (McGill, 1954; Jakulin and Bratko, 2003b). However, it has long been recognized that information theory measures, and many others, generally cannot be computed analytically for all possible subsets of dependent variables. As such, researchers have developed methods that can calculate the presence of nonlinear and high-dimensional dependencies efficiently and reasonably.

Thus, instead of directly calculating the five-way interaction terms, which are computationally expensive, FJMI (Tang et al., 2019) took into account two- through five-way interactions between features and the class variable to capture interactions. The approach is based on the fact that five-dimensional joint mutual information can be decomposed into a sum of two- through five-way interactions, which is easier to compute.

Shishkin et al. (Shishkin et al., 2016b) proposed the CMICOT method, which uses conditional mutual information (CMI) to identify joint interactions between multiple features (more than three). The technique is based on a two-stage greedy search for the approximate solution of high-dimensional CMI and binary representation of features that reduce the dimension of the space of joint distributions, to mitigate the effect of the sample complexity.

Vinh et al. (Vinh et al., 2016) proposed a higher dimensional MI-based feature selection method called RelaxMRMR. To capture higher-order feature interactions, the authors identified the assumptions that can be relaxed for decomposing the full joint mutual information criterion into lower-dimensional MI quantities.

To explicitly treat feature interaction, Zeng et al. (Zeng et al., 2015a) proposed a complementarity-based ranking method called IWFS. The approach is based on interaction weight factors, a variation of three-way interaction that can measure redundancy and complementarity between features.

Based on the link between interaction information and conditional mutual infor-

mation, Cheng et al. (Cheng et al., 2011) proposed a greedy algorithm called CMIFS, which considers not only the competition among features but also the cooperation. This criterion takes account of both redundancy and synergy interactions of features and identifies discriminative features.

El Akadi et al. (El Akadi et al., 2008) proposed an evaluation function called IGFS. It takes into account different features interaction without increasing the computational complexity, and is based on the individual Mutual Information and a compromise (made by the mean of Interaction Gain) between features redundancy and features interaction.

Chow and Huang (Chow and Huang, 2005) combined a pruned Parzen window estimator and the quadratic mutual information for the effective and efficient estimation of high-dimensional mutual information.

With this contribution, Chow and Huang developed a feature selection method called OSF-MI which can identify the salient features and analytically estimate the appropriate feature subsets.

3.4.2 Search Techniques.

Feature selection can be viewed as a search problem, with each state specifying a subset of the relevant features in the search space. An exhaustive method can be used for this purpose in theory but is quite impractical, and in fact, very few feature selection methods use an exhaustive search (Xue et al., 2015). Therefore, heuristic search strategies such as greedy, best-first, and genetic-algorithmic, can be used in a backward elimination or forward selection process for obtaining possible features as a suboptimal solution. However, feature selection problems have a large search space, which is very complex due to feature interaction. To overcome such issues, filter methods that can restrict the solution search space and make the computation more tractable have become essential.

Recently, Singha and Shenoy (Singha and Shenoy, 2018) proposed an adaptive method called SAFE which uses an adaptive 3-way cost function that uses redundancy–complementarity ratio to automatically update the trade-off rule between relevance, redundancy, and complementarity. This approach uses the best-first search strategy, which offers the best compromise solution.

Since it is necessary to balance accuracy and complexity in high-order interactions, Tang et al. (Tang et al., 2018) presented a method called IMFS-FD to obtain a set of features that preserves k-way important interactions but does not intend to interpret all possible interactions reducing the search space.

Mohammadi et al. (Mohammadi et al., 2017) implemented the feature grouping based on multivariate mutual information (FGMMI), which discovers hidden relations between more than two features at the same time. This method aims to construct groups by using the k-means algorithm on a computed MI matrix which divides data into clusters and finally computes MMI for all of the features in each group to select each group's feature having the maximum relevance.

Peng and Liu (Peng, 2016) proposed the RJMIM method that employs a forward greedy search strategy to find and select the features with high discriminative power by measuring both the joint mutual information and the interaction information between the features already selected and candidate features.

Zeng et al. (Zeng et al., 2015b) proposed a feature ranking algorithm called NI-WFS. It is based on neighborhood rough sets that can be used to search for interacting features. Since redundant features produce negative influence and interaction features produce positive influence in predicting, this approach first computes the neighborhood mutual information between a feature and the target and then adjusts it by manipulating the interaction weight factor, which can reflect the information of whether a feature is redundant or interactive.

Bennasar et al. (Bennasar et al., 2013) employs feature interaction – a maximum of the minimum criteria to select the feature that has the strongest relevance to the class label and the highest minimum interaction with the already selected. This method called FIM, is based on three-way interaction information using a forward greedy search algorithm to select relevant and non-redundant features.

To identify all possible feature interactions of maximum size, Sui (Sui, 2013) proposed a BIFS method which is constructed by two main processes: forward identification to identify binary interactions and backward selection where irrelevant feature interaction subsets will be deleted from subsets ranked based on information gain per feature (IGFS).

Zhang and Hancock (Zhang and Hancock, 2011) presented a method called DSplus-MII, which utilizes the multidimensional interaction information criterion and dominant sets for feature selection. This approach can consider third or higher-order feature interactions and limits the resulting search space using dominant set clustering, which separates features into clusters in advance.

Zhao and Liu (Zhao and Liu, 2009) proposed the INTERACT method, which finds interacting features based on a feature sorting metric using data consistency. Contrary to an evaluation based on mutual information, the inconsistency measure is monotonic, allowing an efficient search to explore feature interactions.

For a complementary attribute of an already selected attribute to have a much greater probability of being selected, Meyer and Bontempi (Meyer and Bontempi, 2006) have introduced a method called DISR. Its goal function uses symmetrical relevance and considers the net effect of redundancy and complementarity in the search process. They show that a set of attributes can return information on the class variable that is higher than the sum of the informations from each attribute taken individually.

3.4.3 Number of higher-order interactions.

Despite being hard to measure directly, the interaction and the candidate interactions grow exponentially with the number of features (i.e., the number of variables when considering interactions increases by several orders of magnitude), and higher-order interactions have enormous potential for improving the performance of feature selection. This illustrates why the exploration of high-order interactions is a challenge where increasingly efficient methods have been developed to take into account both 3-way, 4-way and 5-way interactions and can possibly extended to the case of full higher-order terms.

Recently, for instance, Wang et al.(Wang et al., 2021) proposed an algorithm called MRMI to explore three-way interactions. Future works include how it can be extended to the case of higher order terms to select strongly relevant and possibly more interactive features.

To retain the features with the greatest complementarity in the selected feature subset during the progress of feature selection, Li et al.(Li et al., 2020a) proposed a new algorithm, FS-RRC, which computes the complementarity score of two features and the class (three-way interactions).

Pawluk et al. (Pawluk et al., 2019a) proposed a feature selection method named IIFS that considers both 3-way and 4-way interactions. Based on interaction information, they prove some theoretical properties of the novel criterion and the possibility that it may be extended to the case of higher-order terms.

Since the dependence among features is related to both redundancy and complementariness, Chen et al. (Chen et al., 2015) proposed a method called RCDFS where the complementary correlation of features is explicitly separated from redundancy. In this approach, a modification item concerning feature complementariness is introduced in the evaluation criterion in order to identify interaction among more than two features.

Vinh et al. (Vinh et al., 2014) introduced GlobalFS, which can automatically select the number of features to be included and can assess high-order feature dependency

via high dimensional mutual information. However, it is only suitable for problems with a small to medium number of features, e.g., several tens.

Wang et al. (Wang et al., 2013) proposed a rule-based feature selection algorithm FRFS for not only identifying and removing irrelevant and redundant features, but also preserving the interactive ones. The method employs the FOIL algorithm with a restriction to generate classification rules to collect the features whose values appear in the antecedents of the rules generated. Then, it eliminates irrelevant and redundant features while considering multi-way feature interactions.

Bontempi and Meyer (Bontempi and Meyer, 2010) presented mIMR, a causal filter criterion based on three-way interaction that aims to select a feature subset where the most informative variables are the ones having both high mutual information with the class and high complementarity with the others.

To detect pairs of relevant variables that act complementarily in predicting the class, Vergara et al.(Vergara and Estévez, 2010) proposed CMIM-2 as an improvement of the CMIM criterion. It maintains the advantages of the original criterion, but it solves the problem of variables that are relevant in pairs, changing the minimum function to the average function.

Chanda et al. (Chanda et al., 2009) proposed an Interaction Mining (IM) approach to capture the multivariate inter-dependencies (synergy and redundancy) among features, so they employ this k-way interaction information to improve a feature subset selection that has significant interactions with the class variable.

Jakulin and Bratko (Jakulin, 2005) introduced interaction information to measure feature interactions and proposed a feature selection method called ICAP which can detect two-way (one feature and the class) and three-way (two features and the class) interactions.

To close this section we would like to mention that using a chronological perspective, we observe how this research topic receives greater attention from researchers since 2005. It can be concluded that Jakulin's work has had a significant influence on the development of methods for feature selection based on higher-order interaction.

In addition, the entire list of 27 algorithms surveyed, sorted by name, including also full name and reference, is shown in Table 3.1.

3.5 Issues and future challenges

Having seen filter methods that are based on feature intercooperation, some issues arise with maybe subtle distinctions that we'd like to point at, signaling future challenges.

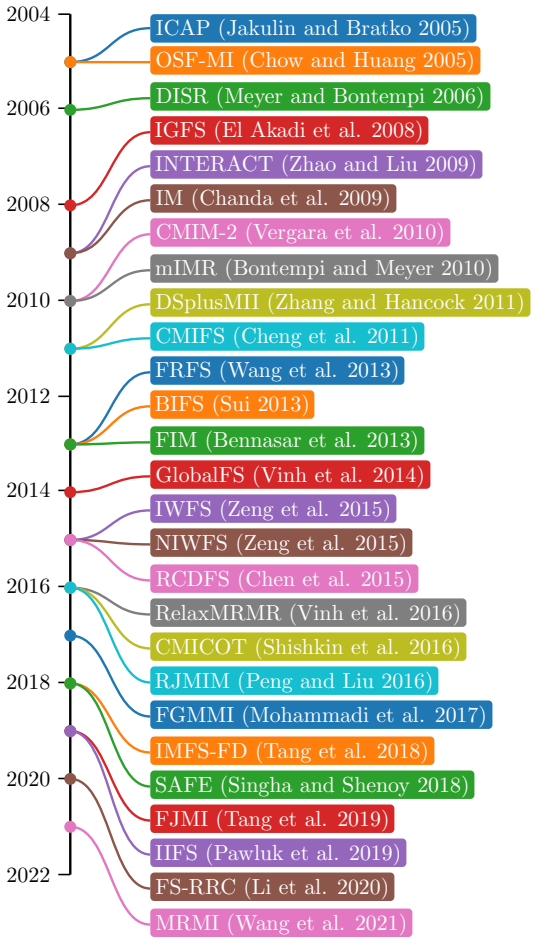


Figure 3.1: Timeline of publications for filter methods based on feature intercooperation. Note that publications in this field are not numerous and represent the results of research initiated in 2005.

3.5.1 Interaction, Synergy, and Complementarity.

In the literature, the terms interaction, synergy, and complementarity are used interchangeably; however, we consider they are not synonymous and have different meanings (Figure 3.2). In this sense, in the most recent study (Wan et al., 2022) an interesting distinction is essentially made between complementarity and interaction, from the point of view of attribute generation into search space; in which the significance of features is assessed through their relevance to the class, redundancy and complementarity with selected features, and interaction with remaining unselected features. Thus, this work complements and extends existing research such that the distinction between interaction, synergy and complementarity is made from the point of view of quantifying multivariate dependencies, and the roles of these variables (i.e. features and/or class).

Table 3.1: Filter Methods based on Feature Intercooperation sorted by name.

Method	Full Name	Reference
BIFS	Binary Interaction based Feature Selection	(Sui, 2013)
CMICOT	Conditional Mutual Information with Complementary and Opposing Teams	(Shishkin et al., 2016b)
CMIFS	Conditional Mutual Information Feature Selection	(Cheng et al., 2011)
CMIM-2	Conditional Mutual Information Maximization Version 2	(Vergara and Estévez, 2010)
DISR	Double Input Symmetrical Relevance	(Meyer and Bontempi, 2006)
DSplusMII	DSplusMII	(Zhang and Hancock, 2011)
FGMMI	Feature Grouping based on Multivariate Mutual Information	(Mohammadi et al., 2017)
FIM	Feature Interaction Maximisation	(Bennasar et al., 2013)
FJMI	Five-way Joint Mutual Information	(Tang et al., 2019)
FRFS	FOIL Rule based Feature Subset Selection	(Wang et al., 2013)
FS-RRC	Feature Selection based on relevance, redundancy and complementarity	(Li et al., 2020a)
GlobalFS	Global Feature Selection	(Vinh et al., 2014)
ICAP/IC	Interaction Capture	(Jakulin, 2005)
IGFS	Interaction Gain for Feature Selection	(El Akadi et al., 2008)
IIFS	Interaction Information Feature Selection	(Pawluk et al., 2019a)
IM	Interaction Mining	(Chanda et al., 2009)
IMFS-FD	Interaction-based Feature Selection using Factorial Design	(Tang et al., 2018)
INTERACT	INTERACT	(Zhao and Liu, 2009)
IWFS	Interaction Weight based Feature Selection	(Zeng et al., 2015a)
mIMR	min-Interaction Max-Relevance	(Bontempi and Meyer, 2010)
MRMI	Max-Relevance Max-Interaction	(Wang et al., 2021)
NIWFS	Neighborhood Interaction Weight based Feature Selection	(Zeng et al., 2015b)
OFS-MI	Optimal Feature Selection using Mutual Information	(Chow and Huang, 2005)
RCDFS	Redundancy-Complementariness Dispersion Feature Selection	(Chen et al., 2015)
RelaxMRMR	RelaxMRMR	(Vinh et al., 2016)
RJMIM	RJMIM	(Peng, 2016)
SAFE	Self-Adaptive Feature Evaluation	(Singha and Shenoy, 2018)

In essence, *interaction* is a measure of dependence between 2 or more variables and can therefore be understood as a nonlinear generalization of correlation. This implies that it can be used to capture a two-way dependence as a minimum order (number of features) in which the class can be included or not.

Definition 3. *There exists **interaction** among variables $X_1, X_2, ..., X_n$ whenever their multivariate symmetrical uncertainty is positive, that is, $MSU(X_1, X_2, ..., X_n) > 0$.*

Now let's consider the question of using interaction to measure the amount of information provided by 2 or more attributes together about the class. In this case, we are talking about *intercooperation* (i.e., multi-way interaction among 2 or more features and the target).

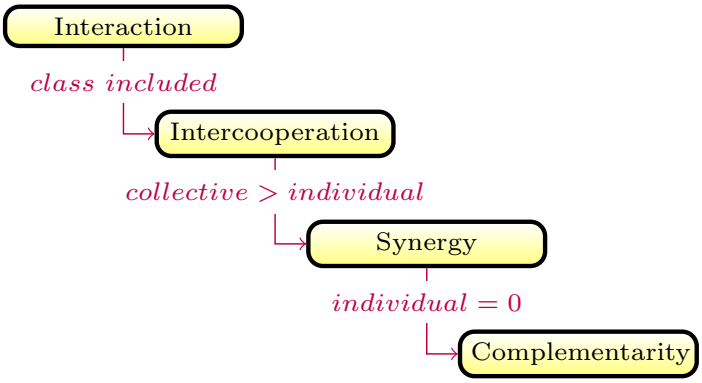


Figure 3.2: Conceptual relationship between terms that differentiates them.

Definition 4. *There exists* ***intercooperation*** *among features* $F_1, F_2, ..., F_n$ *about the class* C *whenever* $MSU(F_1, F_2, ..., F_n, C) - MSU(F_1, F_2, ..., F_n) > 0$.

In this sense, the *synergy* term means that the intercooperation among features provides more information about the class label as a whole than the sum of the individual contributions.

Definition 5. *There exists* ***synergy*** *among features* $F_1, F_2, ..., F_n$ *about the class* C *whenever* $MSU(F_1, F_2, ..., F_n, C) - MSU(F_1, F_2, ..., F_n) > SU(F_1, C) + SU(F_2, C) + \cdots + SU(F_n, C)$.

Following the definition above, *complementarity* occurs when attributes individually do not appear to contain any information about the class and can only contribute in combination with others.

Definition 6. *There exists* ***complementarity*** *among features* $F_1, F_2, ..., F_n$ *about the class* C *whenever* $[MSU(F_1, F_2, ..., F_n, C) - MSU(F_1, F_2, ..., F_n) > 0] \wedge SU(F_i, C) = 0, \forall i \in \{1, ..., n\}$.

3.5.2 Filter method categorization with respect to criterion function scope.

Feature measure or evaluation criterion plays an important role in feature selection, which forms the basis of feature selection (Liu and Motoda, 2012).

Given a target variable $\mathcal{C}$ and $\mathcal{F}$ an n dimensional feature set, where $f_i \in \mathcal{F}$ is used for representing its elements. Let $J(\mathcal{X})$ be a criterion function that evaluates a feature subset $\mathcal{X} \subset \mathcal{F}$. Then the feature selection can be formulated as the problem of finding

an optimal subset of features $\mathcal{S}_{opt}$ for which

$$J(\mathcal{S}_{opt}) = \max_{\mathcal{X} \subset \mathcal{F}} J(\mathcal{X}). \tag{3.6}$$

Consider $\mathcal{S}$ as the subset of currently selected features and f_i as a candidate feature to be added to or deleted from $\mathcal{S}$. Based on the criterion function scope, filter selection methods may roughly be divided into:

3.5.2.1 1st generation filter methods.

They can only measure attributes' relevance according to the amount of individual information contained with respect to the class. These methods are the simplest since their criterion function is defined as:

$$J(f_i) = \mathit{IndividualRelevance}(f_i, \mathcal{C}). \tag{3.7}$$

3.5.2.2 2nd generation filter methods.

While the individual evaluation is incapable of removing redundant features because redundant features are likely to have a similar amount of information, second-generation filter methods can handle feature redundancy with feature relevance through the criterion function:

$$\begin{aligned} J(f_i) \quad = \quad & \mathit{IndividualRelevance}(f_i; \mathcal{C}) \\ & - \mathit{Redundance}(f_i; \mathcal{S}). \end{aligned} \tag{3.8}$$

3.5.2.3 3rd generation filter methods.

A clear limitation of previous approaches is that they neglect a feature that appears to be irrelevant or weakly relevant to the class individually, but when it is combined with other features, it may highly correlate to the class. A concept that was recently considered is intercooperativeness, in which a set of two or more features cooperate to provide information about the target concept:

$$\begin{aligned} J(f_i) = \; & \mathit{IndividualRelevance}(f_i; \mathcal{C}) \\ & - \mathit{Redundance}(f_i; \mathcal{S}) \\ & + \mathit{Intercooperation}(\{f_i, .., f_j\}; \mathcal{S}; \mathcal{C}). \end{aligned} \tag{3.9}$$

3.5.3 Cooperativeness and Exclusive Cooperativeness.

As shown in (Yu and Liu, 2004), finer classification of attribute types might contribute to exploring novel attribute selection strategies, so we propose a conceptual subdivision of attributes into cooperativeness and exclusive cooperativeness according to either independence level or absolute dependence on other attributes in order to provide information. Namely, a cooperative attribute provides information about the class individually and in cooperation with other attributes, while an exclusively cooperative attribute only becomes relevant in the context of others.

3.5.4 Simultaneous Evaluation and Evaluation by Phases.

In (Yu and Liu, 2004), existing approaches to relevance and redundancy were studied. They defined a traditional approach as one that implicitly manages redundancy of attributes with their relevance (i.e., simultaneous evaluation) and proposed another approach in which redundant attributes are explicitly identified for their elimination (i.e., evaluation by phases).

In this regard, when designing a third-generation filter method, we should consider possible cooperation between attributes and the impacts on the scalability and stability resulting from simultaneous evaluation. Thus, scalability is the sensitivity of the computational performance of the feature selection method to data scale, and stability is the sensitivity of feature selection results to training set variations.

Novel feature selection methods need to be developed, in which the evaluation by phases is considered. This approach, which decouples individual relevance analysis, redundancy analysis, and intercooperation analysis, offers alternatives for search space reduction.

3.5.5 Multivariate Dualist Measures.

In (Timme et al., 2014), an interesting perspective was studied: measures that treat all variables equally and measures that treat the class separately from the group of attributes. We refer to the latter ones as multivariate dualist measures. Although this type of measure has been successfully applied to various fields (Lizier et al., 2018), to the best of our knowledge, the use of dualistic multivariate measures for the selection of attributes has not yet been implemented (Yu et al., 2018). Hence, feature selection based on multivariate dualist measures is an interesting possibility.

3.5.6 Maximum Intercooperation Order.

Previous works on real data sets show that the inclusion of high-order dependencies can improve feature selection based on mutual information (Vinh et al., 2016).

However, the number k ($k = 2, 3, 4, 5, m$) of interaction terms is generally determined by expert information, amount of data, degree of error, high-dimensionality assumptions, or some technical considerations such as scalability and/or computation time. As the number of candidate interactions increases exponentially with the number of attributes, it is worth investigating high-order interactions to achieve a balance between accuracy and complexity.

3.5.7 Intercooperation Over/Under Estimation.

Although third generation filter approach overcomes some of the drawbacks of previous generations, it has to deal with new issues. Thus, possible overestimation or underestimation should be considered in the quantification of synergistic information as shown in (Griffith and Koch, 2014). The detection of intercooperativeness itself is a challenge, and therefore, its precise measurement produces a greater challenge.

3.5.8 Redundancy and/xor Synergy.

Many different groups have developed multivariate measures in use today and differ in subtle but significant ways. Thus, a crucial topic related to multivariate information measures is understanding the relationship and meaning of synergy and redundancy. Some authors argue that redundancy and a synergy component can exist simultaneously, whereas others argue that synergy and redundancy are mutually exclusive qualities (Timme et al., 2014).

From the viewpoint of feature selection, the distinction between synergy and redundancy is essential; therefore, their effects are still an open question.

3.5.8.1 Inter-feature redundancy term and complementarity effects.

An interpretation of the objective function of known methods as approximations of a target objective function is proposed in (Macedo et al., 2019).

In the same paper it is verified that a redundancy consisting of the level of association between the candidate attribute and the previously selected attributes is called inter-feature redundancy. Such redundancy is important, for instance, to avoid later problems of collinearity. Furthermore, feature selection methods that include inter-

feature and class-relevant redundancy terms take into account the complementarity expressed as the contribution of a candidate feature to the explanation of the class when taken together with already selected features.

3.5.8.2 Evaluating interaction from the addition of features.

Given a selected feature subset S_j consisting of j variables, suppose we increase the number of variables to k achieving subset S_k so that $S_j \subset S_k$.

If $MSU(S_j) < MSU(S_k)$ the addition of variables has caused a *gain in multiple correlation*, and we can say that the added variables $S_k - S_j$ interact positively with S_j. In the opposite case, if $MSU(S_j) > MSU(S_k)$ we can say that the added variables $S_k - S_j$ interact negatively with S_j.

A proposal of formal definition for interaction in (Gómez-Guerrero et al., 2022) is in terms of k-way interaction on top of j variables: It is the minimum gain in multiple correlation over all possible choices of j-variable subsets S_j within S_k. Note that from a combinatorics point of view, there are $C(k, j)$ possible such subsets.

The proposed definition covers general and complex cases, but it also accommodates the already known classical statistics cases of interaction on a numeric response, occurring in multiple regression and analysis of variance.

3.5.8.3 Intercooperation via Game Theory.

In recent years, other approaches have been investigated to overcome the limitation associated to traditional information-theory-based measures. One of these approaches that have gained popularity is Game Theory (GT).

In GT, the different scenarios are mathematically assessed so that the success of an individual decision depends on the decision choices of others (Von Neumann and Morgenstern, 1947). Azam and You (Azam and Yao, 2011) propose to use GT in feature selection to deal with high imbalance situations in text categorization. Sun et al. (Sun et al., 2012) introduce a cooperative game-theory-based framework to identify the power of each feature according to intricate and intrinsic interrelations among features. Afghagh et al. (Afghah et al., 2018) propose a novel information-theoretic predictive modeling technique based on the idea of coalition game theory for feature selection.

Within GT, the Shapley Value (SV) has been used for feature selection by Chu and Chan (Chu and Chan, 2020). In this work, the SV is decomposed into high-order interaction components to measure the different interaction contributions among features. Bimonte and Senatore (Bimonte and Senatore, 2022) use the SV to construct

the weighted contribution for each feature to allow the selection of features that have explanatory value.

Summarizing, GT, in general, and SV, in particular, are useful approaches to identify the cooperation among features.

3.5.8.4 Feature Selection and/or Deep Learning.

Deep Learning (DL) (LeCun et al., 2015) is an advanced sub-field of Machine Learning that simplifies the modeling of various complex concepts and relationships using multiple levels of representation. DL is distinct from feature selection as DL leverages deep neural networks structures to learn new feature representations while feature selection directly finds relevant features from the original features, thus yielding more readable and interpretable results (Li et al., 2017).

Although DL techniques for attribute selection have shown good results, we believe that more attention should be paid to the importance of attributes and the interpretability of machine learning models, since the most accurate estimates are not always sufficient to solve a data problem.

On the other hand, several studies show that the use of attributes filtered by traditional attribute selection methods and their use as input in a deep generative model outperforms state-of-the-art approaches. Therefore, the study of the effects of intercooperation-based attribute selection methods in a deep generative predictive model are still an open question.

3.6 Conclusions

Feature selection plays an important role in knowledge discovery. It is an effective technique in dealing with dimensionality reduction, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. Therefore, feature selection is active research in the fields of data mining and machine learning. Over the past decade, most research in filter methods has emphasized the use of feature intercooperation to assist in the feature subset selection process. In this chapter, we have surveyed 27 filter feature selection methods that adopt this approach, covering important gaps in the field. In addition, the concepts of relevance, redundancy and intercooperativeness are defined and quantified through information theory measures. Finally, the most significant issues and challenges of filter methods based on feature intercooperation are described, identifying the future research directions in this area.

Chapter 4

PART_FS: A feature selection method based on partitioning and intercooperation[1].

4.1 Introduction.

4.2 Background.

4.3 Problem statement.

4.4 Proposed method.

4.5 Experiments.

4.6 Results.

4.7 Conclusion.

[1]This content is under revision. We will resend the manuscript to you when the new version of the content will be ready. This content is only for the purposes of closing the procedure.

Chapter 5

Conclusions and Future Directions

The advent of Big Data, and specially the advent of datasets with high dimensionality, has brought an important necessity to identify the relevant features of the data. In this scenario, the importance of feature selection is beyond doubt and different methods have been developed, although researchers do not agree on which one is the best method for any given setting (Bolón-Canedo and Alonso-Betanzos, 2018).

In this work, first, we introduce the Multivariate Symmetrical Uncertainty (MSU) measure, as an extension of the Symmetrical Uncertainty (SU) to the multivariate case. In order to evaluate the proposal, several experiments on synthetic datasets are conducted. Results confirm that MSU is a reliable multivariate correlation measure for nominal variables, with promising properties, capable of detecting linear and non-linear dependencies or interactions.

We have also provided a study about the use of feature intercooperation to assist in the feature subset selection process. We have surveyed 27 filter feature selection methods that adopt this approach, covering important gaps in the field of state-of-the-art methods, an issue that has not received much consideration in the literature.

And, finally, a novel feature selection approach based on feature search space partition and features intercooperation named PART_FS is proposed. PART_FS is particularly versatile framework for high-dimensional data of a complex nature. In this sense, we compare the performance of PART_FS on simulated scenarios and real datasets with several recent feature selection methods in combinations with different classifiers. The results show that the proposed method based on partition and intercooperation outperforms the comparison methods and excels in a variety of problems with different characteristics.

Nevertheless, feature selection remains and will continue to be an active field that is incessantly rejuvenating itself to answer new challenges (Liu et al., 2010). For instance,

given that MSU accuracy depends on samples that are totally representative, a main drawback of this (sample size based on total representativeness) consists in the fact that the sample size increases with multivariate cardinality. This implies larger sample sizes to achieve a prescribed precision. Besides, PART_FS performance could be further improved by carefully examining the characteristics of the real datasets, modifying the partitioning criterion and optimizing the model parameters accordingly. Future work will be focused on these points.

References

Afghah, F., Razi, A., Soroushmehr, R., Ghanbari, H., and Najarian, K. (2018). Game theoretic approach for systematic feature selection; application in false alarm detection in intensive care units. *Entropy*, 20(3):190. 60

Ahmed, M. U., Rehman, N., Looney, D., Rutkowski, T. M., Kidmose, P., and Mandic, D. P. (2012). Multivariate entropy analysis with data-driven scales. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3901–3904. 8

Almugren, N. and Alshamlan, H. (2019). A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. *IEEE access*, 7:78533–78548. 42

Arias-Michel, R., García-Torres, M., Schaerer, C., and Divina, F. (2016). Feature selection using approximate multivariate markov blankets. In *Hybrid Artificial Intelligent Systems - 11th International Conference, HAIS 2016, Seville, Spain, April 18-20, 2016, Proceedings*, pages 114–125. 9, 10

Avdiyenko, L., Bertschinger, N., and Jost, J. (2015). Adaptive information-theoretical feature selection for pattern classification. In *Computational Intelligence. Studies in Computational Intelligence*, volume 577, pages 279–294. 9

Azam, N. and Yao, J. (2011). Incorporating game theory in feature selection for text categorization. In *International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing*, pages 215–222. Springer. 60

Bag, S., Kumar, S., Awasthi, A., and Tiwari, M. K. (2019a). A noise correction-based approach to support a recommender system in a highly sparse rating environment. *Decision Support Systems*. 8

Bag, S., Tiwari, M. K., and Chan, F. T. S. (2019b). Predicting the consumer's purchase

intention of durable goods: An attribute-level analysis. *Journal of Business Research*, 94:408–419. 8

Ball, K. R., Granta, C., Mundy, W. R., and Shafer, T. J. (2017). A multivariate extension of mutual information for growing neural networks. *Neural Netowrks*, 95:29–43. 9

Bell, A. J. (2003). The co-information lattice. In *Proceedings of the Fith International Workshop on Independent Component Analysis and Blind Signal Separation*, pages 921–926. 10

Bell, D. A. and Wang, H. (2000). A formalism for relevance and its application in feature subset selection. *Machine learning*, 41(2):175–195. 47

Bennasar, M., Hicks, Y., and Setchi, R. (2015). Feature selection using joint mutual information maximisation. *Expert Systems with Applications*, 42(22):8520 – 8532. 9

Bennasar, M., Setchi, R., and Hicks, Y. (2013). Feature interaction maximisation. *Pattern Recognition Letters*, 34(14):1630–1635. 51, 55

Bimonte, G. and Senatore, L. (2022). Shapley value in partition function form games: New research perspectives for features selection. In *Methods and Applications in Fluorescence*, pages 103–108. Springer. 60

Blum, A. L. and Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1):245 – 271. Relevance. 42

Bolón-Canedo, V. and Alonso-Betanzos, A. (2018). *Recent advances in ensembles for feature selection*, volume 147. Springer. 63, 92

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2016). Feature selection for high-dimensional data. *Progress in Artificial Intelligence*, 5(2):65–75. 42

Bontempi, G. and Meyer, P. E. (2010). Causal filter selection in microarray data. In *Proceedings of the 27th international conference on machine learning (icml-10)*, pages 95–102. 53, 55

Brown, G. (2009). A new perspective for information theoretic feature selection. In van Dyk, D. and Welling, M., editors, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 49–56. PMLR. 10

Cai, J., Luo, J., Wang, S., and Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, 300:70–79. 48

Caruana, R. and Freitag, D. (1994). How useful is relevance? *FOCUS*, 14(8):2. 47

Chan, C., Al-Bashabsheh, A., and Zhou, Q. (2018). Change of multivariate mutual information: From local to global. *IEEE Transactions on Information Theory*, 64(1):57–76. 9

Chanda, P., Cho, Y.-R., Zhang, A., and Ramanathan, M. (2009). Mining of attribute interactions using information theoretic metrics. In *2009 IEEE International Conference on Data Mining Workshops*, pages 350–355. IEEE. 41, 53, 55

Chen, Z., Wu, C., Zhang, Y., Huang, Z., Ran, B., Zhong, M., and Lyu, N. (2015). Feature selection with redundancy-complementariness dispersion. *Knowledge-Based Systems*, 89:203–217. 2, 10, 42, 52, 55, 78

Cheng, g., Qin, Z., Feng, C., Wang, Y., and Li, F. (2011). Conditional mutual information-based feature selection analyzing for synergy and redundancy. *Etri Journal*, 33(2):210–218. 50, 55

Chow, T. W. and Huang, D. (2005). Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information. *IEEE Transactions on Neural networks*, 16(1):213–224. 50, 55

Chu, C. C. F. and Chan, D. P. K. (2020). Feature selection using approximated high-order interaction components of the shapley value for boosted tree classifier. *IEEE Access*, 8:112742–112750. 60

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons. 82, 83, 84

Croux, C. and Dehon, C. (2010). Influence functions of the spearman and kendall correlation measures. *Statistical methods & applications*, 19(4):497–515. 44

Curtiss, J. (1941). On the distribution of the quotient of two chance variables. *The Annals of Mathematical Statistics*, 12(4):409—421. 20

Doquire, G. and Verleysen, M. (2012). A comparison of multivariate mutual information estimators for feature selection. In *ICPRAM 2012 - Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods, Volume 1, Vilamoura, Algarve, Portugal, 6-8 February, 2012*, pages 176–185. 10

Eibe, F., Hall, M. A., and Witten, I. H. (2016). The weka workbench. online appendix for data mining: practical machine learning tools and techniques. In *Morgan Kaufmann*. Morgan Kaufmann Publishers San Francisco, California. 92

El Akadi, A., El Ouardighi, A., and Aboutajdine, D. (2008). A powerful feature selection approach based on mutual information. *International Journal of Computer Science and Network Security*, 8(4):116. 50, 55

Fayyad, U. M. and Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *Ijcai*, volume 93, pages 1022–1029. Citeseer. 84

García-Torres, M., Gómez-Vela, F., Melián-Batista, B., and Moreno-Vega, J. M. (2016). High-dimensional feature selection via feature grouping: A variable neighborhood search approach. *Information Sciences*, 326:102 – 118. 9

Griffith, V. and Koch, C. (2014). Quantifying synergistic mutual information. In *Guided Self-Organization: Inception*, pages 159–190. Springer. 59

Guo, B. and Nixon, M. S. (2009). Gait feature subset selection by mutual information. *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, 39(1):36–46. 9

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182. 2, 42, 43, 78

Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. A. (2008). *Feature extraction: foundations and applications*, volume 207. Springer. 42

Gómez-Guerrero, S., Ortiz, I., Sosa-Cabrera, G., García-Torres, M., and Schaerer, C. E. (2022). Measuring interactions in categorical datasets using multivariate symmetrical uncertainty. *Entropy*, 24(1). 60

Hall, M. A. (1998). *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, Hamilton, New Zealand. 8, 9, 38

Han, T. (1980). Slepian-wolf-cover theorem for network of channels. *Info. and Contr.*, 47(1):67–83. 43

Höeppner, F. and Klawann, F. (2008). *Handbook of Granular Computing*, chapter Systems of Information Granules, pages 187–204. Wiley. 8

Ince, R. A. A. (2017). Measuring multivariate redundant information with pointwise common change in surprisal. *Entropy*, 19(7):1–37. 10

Jakulin, A. (2005). *Machine learning based on attribute interactions*. PhD thesis, Univerza v Ljubljani. 46, 47, 53, 55

Jakulin, A. and Bratko, I. (2003a). Analyzing attribute dependencies. In *European conference on principles of data mining and knowledge discovery*, pages 229–240. Springer. 2, 78

Jakulin, A. and Bratko, I. (2003b). Quantifying and visualizing attribute interactions. *CoRR*, cs.AI/0308002. 10, 42, 49

Jakulin, A. and Bratko, I. (2004). Testing the significance of attribute interactions. In *Proceedings of the twenty-first international conference on Machine learning*, page 52. ACM. 43, 45

Jesus, J., Canuto, A., and Araújo, D. (2017). A feature selection approach based on information theory for classification tasks. In *Artificial Neural Networks and Machine Learning – ICANN 2017*, volume 10614 of *Lecture Notes in Computer Science*, pages 359–367. Springer-Verlag. 9

John, G. H., Kohavi, R., and Pfleger, K. (1994). Irrelevant features and the subset selection problem. In *Machine learning proceedings 1994*, pages 121–129. Elsevier. 27

Kashef, S. and Nezamabadi-pour, H. (2019). A label-specific multi-label feature selection algorithm based on the pareto dominance concept. *Pattern Recognition*, 88:654–667. 8

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2):273–324. 42

Kojadinovic, I. (2005). Relevance measures for subset variable selection in regression problems based on k-additive mutual information. *Computational Statistics & Data Analysis*, 49(4):1205–1227. 10, 43

Koller, D. and Sahami, M. (1996). Toward optimal feature selection. Technical report, Stanford InfoLab. 47

Kononenko, I. (1995). On biases in estimating multi-valued attributes. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, pages 1034–1040. Morgan Kaufmann Publishers Inc. 26, 27

Kraskov, A., Stögbauer, H., Andrzejak, R. G., and Grassberger, P. (2003). Hierarchical clustering based on mutual information. *Computing Research Repository*, q-bio.QM/0311039. 84

Lavangnananda, K. and Chattanachot, S. (2017). Study of discretization methods in classification. In *2017 9th International Conference on Knowledge and Smart Technology (KST)*, pages 50–55. IEEE. 44

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444. 61

Li, C., Luo, X., Qi, Y., Gao, Z., and Lin, X. (2020a). A new feature selection algorithm based on relevance, redundancy and complementarity. *Computers in Biology and Medicine*, 119:103667. 52, 55

Li, C., Luo, X., Qi, Y., Gao, Z., and Lin, X. (2020b). A new feature selection algorithm based on relevance, redundancy and complementarity. *Computers in biology and medicine*, 119:103667. 91

Li, F., Zhang, Z., and Jin, C. (2016). Feature selection with partition differentiation entropy for large-scale data sets. *Information Sciences*, 329:690 – 700. Special issue on Discovery Science. 9

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., and Liu, H. (2017). Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50(6):1–45. 61

Liu, H. and Motoda, H. (2012). *Feature selection for knowledge discovery and data mining*, volume 454. Springer Science & Business Media. 42, 44, 56

Liu, H., Motoda, H., Setiono, R., and Zhao, Z. (2010). Feature selection: An ever evolving frontier in data mining. In *Feature selection in data mining*, pages 4–13. PMLR. 2, 42, 63, 94

Liu, H. and Zhao, Z. (2012). Manipulating data and dimension reduction methods: Feature selection. *Computational Complexity: Theory, Techniques, and Applications*, pages 1790–1800. 42

Lizier, J., Bertschinger, N., Jost, J., and Wibral, M. (2018). Information decomposition of target effects from multi-source interactions: Perspectives on previous, current and future work. 58

Macedo, F., Oliveira, M. R., Pacheco, A., and Valadas, R. (2019). Theoretical foundations of forward feature selection methods based on mutual information. *Neurocomputing*, 325:67–89. 42, 59

Markelle Kelly, Rachel Longjohn, K. N. (2023). The uci machine learning repository. 91

McGill, W. J. (1954). Multivariate information transmission. *Psychometrika*, 19(2):97–116. 10, 13, 43, 46, 49, 85

Méndez, J. R., Cotos-Yañez, T. R., and Ruano-Ordás, D. (2019). A new semantic-based feature selection method for spam filtering. *Applied Soft Computing*, 76:89–104. 8

Meyer, P. E. and Bontempi, G. (2006). On the use of variable complementarity for feature selection in cancer classification. In *Workshops on applications of evolutionary computation*, pages 91–102. Springer. 52, 55

Mohammadi, S., Mirvaziri, H., and Ahsaee, M. G. (2017). Multivariate correlation coefficient and mutual information-based feature selection in intrusion detection. *Information Security Journal: A Global Perspective*, 26(5):229–239. 10, 51, 55

Ni, L., Cao, J., and Wang, R. (2013). *Time-Dependent Multivariate Multiscale Entropy Based Analysis on Brain Consciousness Diagnosis. In: Advances in Brain Inspired Cognitive Systems, Lecture Notes in Computer Science*, volume 7888. Springer-Verlag, Berlin, Germany. 8

Palma-Mendoza, R.-J., de Marcos, L., Rodriguez, D., and Alonso-Betanzos, A. (2018). Distributed correlation-based feature selection in spark. *Information Sciences*. 8

Pawluk, M., Teisseyre, P., and Mielniczuk, J. (2019a). Information-theoretic feature selection using high-order interactions. *Machine Learning, Optimization, and Data Science*, pages 51–63. 52, 55

Pawluk, M., Teisseyre, P., and Mielniczuk, J. (2019b). Information-theoretic feature selection using high-order interactions. In *Machine Learning, Optimization, and Data Science: 4th International Conference, LOD 2018, Volterra, Italy, September 13-16, 2018, Revised Selected Papers 4*, pages 51–63. Springer. 91

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238. 9

Peng, L. (2016). Rjmim: A new feature selection method based on joint mutual information. *Revista de la Facultad de Ingeniería*, 31(4). 51, 55

Pham, T. H., Ho, T. B., Nguyen, Q. D., Tran, D. H., and Nguyen, V. H. (2012). Multivariate mutual information measures for discovering biological networks. In *2012 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future*, pages 1–6. IEEE. 9

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1988). *Numerical Recipes in C*. Cambridge University Press, Cambridge, UK. 12, 45

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 12, 45, 83

Schouten, B., Cobben, F., and Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 34(1):101–113. 16

Shalizi, C. (2009). Information and interaction among features (notes chapter, Statistics Department, Carnegie-Mellon University). 10

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423. 8, 81, 83

Shin, K., Kuboyama, T., Hashimoto, T., and Shepard, D. (2017). Scwc/slcc: Highly scalable feature selection algorithms. *Information*, 8(4):159. 86

Shishkin, A., Bezzubtseva, A., Drutsa, A., Shishkov, I., Gladkikh, E., Gusev, G., and Serdyukov, P. (2016a). Efficient high-order interaction-aware feature selection based on conditional mutual information. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 4637–4645. The MIT Press. 9

Shishkin, A., Bezzubtseva, A., Drutsa, A., Shishkov, I., Gladkikh, E., Gusev, G., and Serdyukov, P. (2016b). Efficient high-order interaction-aware feature selection based on conditional mutual information. In *Advances in neural information processing systems*, pages 4637–4645. 49, 55

Shlomo, N., Skinner, C., and Schouten, B. (2012). Estimation of an indicator of the representativeness of survey response. *Journal of Statistical Planning and Inference*, 142:201–211. 16

Singh, P. K. (2018). m-polar fuzzy graph representation of concept lattice. *Engineering Applications of Artificial Intelligence*, 67:52–62. 8

Singh, P. K., Cherukuri, A. K., and Li, J. (2017). Concepts reduction in formal concept analysis with fuzzy setting using shannon entropy. *International Journal of Machine Learning and Cybernetics*, 8:179–189. 8

Singh, P. K. and Gani, A. (2015). Fuzzy concept lattice reduction using shannon entropy and huffman. *Journal of Applied Non-Classical Logics*, 25(2):101–119. 17

Singha, S. and Shenoy, P. P. (2018). An adaptive heuristic for feature selection based on complementarity. *Machine Learning*, 107(12):2027–2071. 10, 50, 55, 91

Sosa-Cabrera, G., García-Torres, M., Gómez-Guerrero, S., Schaerer, C. E., and Divina, F. (2019). A multivariate approach to the symmetrical uncertainty measure: Application to feature selection problem. *Information Sciences*, 494:1–20. 4, 46, 47, 85, 93

Sosa-Cabrera, G., Gómez-Guerrero, S., García-Torres, M., and Schaerer, C. E. (2023). Feature selection: a perspective on inter-attribute cooperation. *International Journal of Data Science and Analytics*, pages 1–13. 5, 90, 91, 93

Studený, M. and Vejnarová, J. (1998). The multiinformation function as a tool for measuring stochastic dependence. In *Proceedings of the NATO Advanced Study Institute on Learning in graphical models*, pages 261–297. 10

Sui, B. (2013). *Information gain feature selection based on feature interactions*. PhD thesis, Citeseer. 1, 51, 55, 77

Sun, X., Liu, Y., Li, J., Zhu, J., Chen, H., and Liu, X. (2012). Feature evaluation and selection with cooperative game theory. *Pattern recognition*, 45(8):2992–3002. 60

Tang, X., Dai, Y., Sun, P., and Meng, S. (2018). Interaction-based feature selection using factorial design. *Neurocomputing*, 281:47–54. 50, 55, 90

Tang, X., Dai, Y., and Xiang, Y. (2019). Feature selection based on feature interactions with application to text categorization. *Expert Systems with Applications*, 120:207–216. 49, 55, 90

Thompson, S. K. (1987). Sample size for estimating multinomial proportions. *The American Statistician*, 41(1):42–46. 23

Thrun, S. B., Bala, J., Bloedorn, E., Bratko, I., Cestnik, B., Cheng, J., Jonng, K. D., Dzeroski, S., Fahlman, S. E., Fisher, D., Hamann, R., Kaufman, K., Keller, S., Kononenko, I., Kreuziger, J., Michalski, R. S., Mitchell, T., Pachowicz, P., Reich, Y., Vafaie, H., de Welde, W. V., Wenzel, W., Wnek, J., and Zhang., J. (1992). The MONK's problems: A performance comparison of different learning algorithms. Technical report, Carnegie Mellon University. 27

Timme, N., Alford, W., Flecker, B., and Beggs, J. M. (2014). Synergy, redundancy, and multivariate information measures: an experimentalist's perspective. *Journal of computational neuroscience*, 36(2):119–140. 58, 59

Tsujishita, T. (1995). On triple mutual information. *Advances in applied mathematics*, 16(3):269–274. 43

Ullah, A., Qamar, U., Khan, F. H., and Bashir, S. (2017). Dimensionality reduction approaches and evolving challenges in high dimensional data. In *Proceedings of the 1st International Conference on Internet of Things and Machine Learning*, page 67. ACM. 48

Vergara, J. R. and Estévez, P. A. (2010). Cmim-2: an enhanced conditional mutual information maximization criterion for feature selection. *Journal of Applied Computer Science Methods*, 2. 53, 55

Vergara, J. R. and Estévez, P. A. (2014). A review of feature selection methods based on mutual information. *Neural computing and applications*, 24(1):175–186. 43, 47

Vinh, N. X., Chan, J., and Bailey, J. (2014). Reconsidering mutual information based feature selection: A statistical significance view. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*. 52, 55

Vinh, N. X., Zhou, S., Chan, J., and Bailey, J. (2016). Can high-order dependencies improve mutual information based feature selection? *Pattern Recognition*, 53:46–58. 49, 55, 59, 90

Von Neumann, J. and Morgenstern, O. (1947). *Theory of games and economic behavior, 2nd rev.* Princeton university press. 60

Wan, J., Chen, H., Li, T., Huang, W., Li, M., and Luo, C. (2022). R2ci: Information theoretic-guided feature selection with multiple correlations. *Pattern Recognition*, 127:108603. 42, 43, 54

Wang, G., Song, Q., Xu, B., and Zhou, Y. (2013). Selecting feature subset for high dimensional data via the propositional foil rules. *Pattern Recognition*, 46(1):199–214. 47, 53, 55

Wang, L., Jiang, S., and Jiang, S. (2021). A feature selection method via analysis of relevance, redundancy, and interaction. *Expert Systems with Applications*, 183:115365. 52, 55

Watanabe, S. (1960). Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, 4(1):66–82. 10, 13, 85

Xue, B., Zhang, M., Browne, W. N., and Yao, X. (2015). A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on evolutionary computation*, 20(4):606–626. 43, 50

Yang, H. H. and Moody, J. (1999). Feature selection based on joint mutual information. In *Proceedings of International ICSC Symposium on Advances in Intelligent Data Analysis*, pages 22–25. 9

Yao, G., Hu, X., and Wang, G. (2022). A novel ensemble feature selection method by integrating multiple ranking information combined with an svm ensemble model for enterprise credit risk prediction in the supply chain. *Expert Systems with Applications*, 200:117002. 41

Yeung, R. W. (1991). A new outlook on shannon's information measures. *IEEE transactions on information theory*, 37(3):466–474. 43

Yu, L. and Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of machine learning research*, 5(Oct):1205–1224. 9, 47, 58

Yu, S., Giraldo, L. G. S., Jenssen, R., and Principe, J. C. (2018). Multivariate extension of matrix-based renyi's {\alpha}-order entropy functional. *arXiv preprint arXiv:1808.07912*. 58

Zeng, Z., Zhang, H., Zhang, R., and Yin, C. (2015a). A novel feature selection method considering feature interaction. *Pattern Recognition*, 48(8):2656–2666. 2, 42, 49, 55, 78, 90

Zeng, Z., Zhang, H., Zhang, R., Zhang, Y., et al. (2015b). A mixed feature selection method considering interaction. *Mathematical Problems in Engineering*. 2, 51, 55, 78

Zhang, Y., Li, S., Wang, T., and Zhang, Z. (2013). Divergence-based feature selection for separate classes. *Neurocomputing*, 101:32–42. 86

Zhang, Z. and Hancock, E. R. (2011). A graph-based approach to feature selection. In *International Workshop on Graph-Based Representations in Pattern Recognition*, pages 205–214. Springer. 51, 55

Zhao, Z. and Liu, H. (2009). Searching for interacting features in subset selection. *Intelligent Data Analysis*, 13(2):207–228. 51, 55, 90

Zhong, N., Dong, J., and Ohsuga, S. (2004). Using rough sets with heuristics for feature selection. *Journal of Intelligent Information Systems*, 16:199–214. 42

Appendix A

A Summary in Spanish

A.1 Introducción

En las tareas de clasificación, un atributo (i.e., una variable independiente) se considera relevante, irrelevante o redundante en función a la información que contiene acerca del concepto objetivo o clase (i.e., la variable dependiente).

La selección de atributos se define como el método de encontrar un conjunto mínimo de atributos relevantes con el objetivo de minimizar el error en la proceso de clasificación con respecto a una determinada clase.

En este sentido, la selección de atributos se ha convertido en el punto clave de gran parte de la investigación en áreas en las que intervienen conjuntos de datos de alta dimensión. Entre estas áreas se encuentran el procesamiento de textos, la expresión genética y la química combinatoria (Sui, 2013).

Un método de selección de atributos tiene tres componentes: la definición del criterio de evaluación (e.g., la relevancia de los atributos), la estimación del criterio de evaluación (i.e., la medida) y las estrategias de búsqueda para la generación de subconjuntos de atributos candidatos.

Con relación a las medidas de evaluación, se han propuesto varios criterios para evaluar los atributos y determinar su importancia. Cabe destacar que basado en los criterios de evaluación, los metodos de selección de atributos pueden ser divididos en envolvente (wrapper), filtro (filter) y embebido (embedded).

En los métodos del tipo filtro, la evaluación del subconjunto de atributos se lleva a cabo por medio de la valoración de las propiedades intrínsecas del dato, tales como la distancia, la consistencia, la entropía y la correlación.

Esta estrategia no considera ninguna relación con el algoritmo de aprendizaje por lo que son mucho más eficientes en términos de recursos computacionales ya son eje-

cutados como una etapa previa y denominada de preprocesamiento.

Aunque la literatura ofrece una amplia y variada gama de métodos de selección de atributos de tipo filtro, en la mayoría de los casos, se trata únicamente de la identificación de atributos irrelevantes y redundantes, donde un aspecto importante que habitualmente se descuida es la complementariedad de atributos (Guyon and Elisseeff, 2003; Chen et al., 2015) (también conocida como sinergia (Zeng et al., 2015a) o interacción (Jakulin and Bratko, 2003a)).

Los atributos que interactúan son aquellos que parecen ser irrelevantes o poco relevantes para la clase cuando son considerados individualmente, pero que cuando se combinan con otros atributos, pueden tener una alta correlación con la clase (Zeng et al., 2015b).

Una motivación para el desarrollo de esta tesis, es que la interacción de atributos ha recibido una atención considerable en los últimos tiempos y despierta cada vez más la atención de los investigadores (Zeng et al., 2015b,a).

Entre otras razones, es que a lo largo de las décadas, los métodos de selección de atributos han evolucionado desde los simples algoritmos de clasificación de relevancia univariante, pasando por los de compensación de relevancia-redundancia; hasta los más sofisticados enfoques basados en las dependencias multivariantes en los últimos años.

Esta tendencia a capturar la dependencia multivariante tiene como objetivo obtener información única sobre la clase a partir de lo que en este estudio se define como intercooperación entre atributos.

Es por ello, que se pretende en esta tesis proponer formas de detectar, medir e identificar cuáles son las asociaciones entre atributos que aportan colectivamente información única acerca de la variable explicada o clase del caso y sus implicancias en la busqueda de un subconjunto mínimo de atributos relevantes.

A.1.1 Objetivos

A.1.1.1 Objetivo General

Examinar la dependencia multivariante categórica mediante su detección, cuantificación y caracterización orientado al proceso de selección de atributos aplicado en la tarea de clasificación de la minería de datos.

A.1.1.2 Objetivos Específicos

- Definir y explorar una medida de dependencia multivariable basado en la teoría de la información como lo es la Incertidumbre Simétrica Multivariada (MSU,

Multivariate Symmetrical Uncertainty).

- Determinar los límites de las medidas de información multivariables en las estrategias de búsqueda más utilizadas en el proceso de Selección de Atributos.
- Establecer y caracterizar las nociones concernientes a la dependencia multivariada en el contexto de la Selección de Atributos.
- Elaborar una revisión sistemática del estado del arte sobre los heurísticos de selección de atributos basados en la detección y/o cuantificación de la dependencia multivariada.
- Idear un heurístico para la selección de atributos mediante el aprovechamiento de la detección de la dependencia multivariante.

A.1.2 Contribuciones de la Tesis

En este trabajo de tesis, se han abordado diferentes aspectos clave relacionado con la dependencia multivariable en el contexto de la selección de atributos. Las aportaciones más significativas de esta tesis son expuestos a continuación:

- Definición y análisis de la Incertidumbre Simétrica Multivariada (MSU, Multivariate Symmetrical Uncertainty) como medida de información de orden superior aplicable al proceso de Selección de Atributos.
- Estudio del rendimiento del MSU bajo densidades de datos con patrones conocidos en la práctica y con conjuntos de datos reales de interés para el país y la región.
- Formulación genérica y caracterización de las nociones referente a la selección de atributos asistido por intercooperación.
- Revisión del estado del arte acerca de los heurísticos de selección de atributos basados en la dependencia multivariable donde se resume las contribuciones de los diferentes enfoques encontrados en la literatura. Además, se presentan los problemas y retos actuales para identificar los métodos más prometedores dado los vacíos específicos del conocimiento en el tema.
- Propuesta de un método novedoso de selección de atributos basado en la partición del espacio de búsqueda de atributos y la intercooperación de los mismos. Este método utiliza KMedoids para la partición en subespacios, además de utilizar

medidas basadas en información y consistencia para abordar la intercooperatividad.

- Desarrollo de un *toolbox* implementado en *PYTHON* para realizar selección de atributos usando medidas de dependencia multivariadas. Este *toolbox* implementa los métodos principales basados en la intercooperación de atributos, además del método propuesto.

En adición, cabe destacar que en la búsqueda de soluciones y mejoras a las diferentes cuestiones planteadas, han surgido nuevas ideas e inquietudes no exploradas con profundidad. Estos temas de trabajo constituyen la semilla para nuevas líneas de investigación a partir de la presente tesis.

A.1.3 Publicaciones

Los capítulos principales de esta propuesta de tesis doctoral se derivan de los siguientes artículos publicados o sometidos en proceso de revisión.

- **Sosa-Cabrera, G.**, Gómez-Guerrero, S., García-Torres, M., & Schaerer, C. E. (2023). *PART_FS: A feature selection method based on partitioning and intercooperation.* Status: In review.

- **Sosa-Cabrera, G.**, Gómez-Guerrero, S., García-Torres, M., & Schaerer, C. E. (2023). *Feature selection: a perspective on inter-attribute cooperation.* International Journal of Data Science and Analytics, 1-13.

- **Sosa-Cabrera, G.**, García-Torres, M., Gómez-Guerrero, S., Schaerer, C. E., & Divina, F. (2019). *A multivariate approach to the symmetrical uncertainty measure: application to feature selection problem.* Information Sciences, 494, 1-20.

Las publicaciones del autor en temas de investigación relacionados incluidos en la presente tesis son:

- **Sosa-Cabrera, G.**, Torres, M. G., Guerrero, S. G., Schaerer, C. E., & Divina, F. (2018). *Understanding a multivariate semi-metric in the search strategies for attributes subset selection.* Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, 6(2).

- Gómez-Guerrero, S., Ortiz, I., **Sosa-Cabrera, G.**, García-Torres, M., & Schaerer, C. E. (2021). *Measuring Interactions in Categorical Datasets Using Multivariate Symmetrical Uncertainty.* Entropy, 24(1), 64.

- Gómez-Guerrero, S., **Sosa-Cabrera, G.**, García-Torres, M., Ortiz-Samudio, I., & Schaerer, C. E. (2021). *Multivariate Symmetrical Uncertainty as a measure for interaction in categorical patterned datasets.* Proceedings of the Entropy 2021: The Scientific Tool of the 21st Century session Information Theory, Probability and Statistics.

- Gómez-Guerrero, S., García-Torres, M., **Sosa-Cabrera, G.**, Sotto-Riveros, E., & Schaerer, C. E. (2021). *Classifying dengue cases using CatPCA in combination with the MSU correlation.* Proceedings of the Entropy 2021: The Scientific Tool of the 21st Century session Entropy in Multidisciplinary Applications.

Las publicaciones del autor en temas de investigación relacionados no incluidos en la presente tesis son:

- **Sosa-Cabrera, G.**, Torres, M. G., Guerrero, S. G., Schaerer, C. E., & Divina, F. (2018). *Effect of Sample Representativeness in Multivariate Symmetrical Uncertainty for Categorical Attributes.* Proceedings of the Third Conference on Business Analytics in Finance and Industry.

- **Sosa-Cabrera, G.**, Torres, M. G., Guerrero, S. G., Schaerer, C. E. (2018). *Is it correlation or interaction?.* En III Encuentro de Investigadores de la Sociedad Científica del Paraguay.

A.2 Fundamentación Teórica

En esta sección llevaremos a cabo una revisión de las nociones de la teoría de la información y de la selección de atributos que son mencionados a lo largo del presente trabajo y cuyo interés radica en que pueden ser utilizados con el objeto de medir la cantidad de información como una reducción de la incertidumbre.

A.2.1 Teoría de la Información

A.2.1.1 Entropía.

La entropía de Shannon (H) (Shannon, 1948) de una variable aleatoria discreta X, con $\{x_1, \ldots, x_n\}$ como posibles valores y la función de probabilidad $P(X)$, es una medida

de la incertidumbre en la predicción del valor de X.

Definition 7. *La entropía $H(X)$ se define como*

$$H(X) := -\sum_i P(x_i) \log_2(P(x_i)), \tag{A.1}$$

donde $P(x_i)$ es la probabilidad de la variable X y la sumatoria se produce sobre todos los valores posibles de X, denotado por x_i. $H(X)$ puede ser también interpretado como una medida de la variedad inherente a X, o la cantidad de información que es necesaria para predecir o describir el resultado de X.

Entropía Conjunta. Para variables independientes (X, Y) con $P(X, Y)$ como distribución conjunta de probabilidad se tiene la entropía conjunta $H(X, Y)$.

Definition 8. *La entropía conjunta $H(X, Y)$ está definida como*

$$H(X, Y) := -\sum_{x \in X} \sum_{y \in Y} P(x, y) \log_2[P(x, y)]. \tag{A.2}$$

Entropía Condicional. La entropía condicional $H(X|Y)$ cuantifica la cantidad de información necesaria para describir el resultado de X dado que el valor de otra variable aleatoria discreta Y es conocido.

Definition 9. *La entropía condicional está definida como*

$$H(X|Y) := -\sum_j \left[P(y_j) \sum_i P(x_i|y_j) \log_2(P(x_i|y_j)) \right],$$

donde $P(y_j)$ es la probabilidad a priori del valor y_j de Y, y $P(x_i|y_j)$ es la probabilidad a posteriori de un valor x_i para la variable X puesto que el valor de la variable Y es y_j.

Propiedades de la Entropía

La entropía de Shannon satisface las siguientes propiedades (Cover and Thomas, 2012):

1. $H(X) \geq 0$. No-negatividad.
2. $H_b(X) = (\log_b a) H_a(X)$. Cambio de base del logaritmo.
3. $H(X|Y) \leq H(X)$. El condicionamiento reduce la entropía siendo $H(X|Y) = H(X)$ si y solamente si X e Y son independientes.

4. $H(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$, como cota superior y llegando a la igualdad si y solamente sí las variables aleatorias X_i son independientes.

5. $H(X) \leq \log(X)$, como cota superior y llegando a la igualdad si X es una variable aleatoria uniforme.

Theorem A.2.1. *(regla de la cadena) (Cover and Thomas, 2012). Dadas dos variables aleatorias X e Y la entropía conjunta está dada por $H(X, Y) = H(X) + H(Y|X)$.*

Una extensión del Teorema A.2.1 para X e Y dado Z está expresado por el corolario

Corollary A.2.2. $H(X, Y|Z) = H(X|Z) + H(Y|X, Z)$.

Una generalización del Teorema A.2.1 está dado por

Theorem A.2.3. *(regla de la cadena general) (Cover and Thomas, 2012). En general, se cumple la regla de la cadena para múltiples variables aleatorias: $H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_{i-1})$.*

A.2.1.2 Ganancia de la Información

Conocido alternativamente como Información Mutua (Shannon, 1948), la Ganancia de la Información ($IG(X|Y)$) (Quinlan, 1993) de una variable X con respecto a otra variable dada Y mide la reducción de la incertidumbre acerca del valor de la variable X cuando el valor de Y es conocido.

Definition 10. *La información ganada de X al conocer Y se define como*

$$IG(X|Y) := H(X) - H(X|Y). \tag{A.3}$$

IG mide cuánto el conocimiento de Y hace que el valor de X sea más fácil de predecir, y por tanto, el mismo puede ser utilizado como una *medida de correlación*.

Observe que como casos extremos son obtenidos

1. si X e Y son independientes, entonces $IG(X|Y) = 0$, y

2. si X e Y son completamente correlacionados entonces $H(X|Y) = 0$ y por tanto $IG(X|Y) = H(X)$.

En general, para variables aleatorias cualesquiera X, Z e Y, $IG(X|Y) > IG(Z|Y)$ significa que conociendo el valor de Y la reducción en la incertidumbre acerca de X es mayor que la reducción en la incertidumbre acerca de Z, si X está más correlacionado a Y que este a Z.

Se puede demostrar que $IG(X;Y)$ es una medida simétrica, lo cual es una conveniente propiedad para una medida entre dos variables ya que el orden entre ellas no altera el resultado de la medición. Es decir, $IG(X;Y) = IG(Y;X)$.

Sin embargo, IG presenta un inconveniente: cuando X y/o Y tiene más valores posibles, ellas aparecen con mayor correlación, por tanto, IG tiende a ser más alto cuando se presentan variables con mayor número de valores posibles. Esto es equivalente a decir que el IG es dependiente de la cardinalidad.

Propiedades de la ganancia de la información

1. $IG(X;Y) \geq 0$.
2. $IG(X;Y) = H(X) - H(X|Y)$.
3. $IG(X;Y) = H(Y) - H(Y|X)$.
4. $IG(X;Y) = H(X) + H(Y) - H(X,Y)$.
5. $IG(X;Y) = IG(Y;X)$ (simetría).
6. $IG(X;X) = H(X)$ (información propia).

Dada la relación con la entropía, para la ganancia de la información se puede establecer el teorema

Theorem A.2.4. *(regla de la cadena) (Cover and Thomas, 2012). Para un conjunto de n variables* $\{X_1, \ldots, X_n\}$ *e* Y *la ganancia de la información está dada por* $IG(X_1, \ldots, X_n; Y) = \sum_{i=1}^{n} IG(X_i; Y|X_1, \ldots, X_{i-1})$.

La unidad de información del IG depende de la base del logaritmo utilizado. En el presente trabajo se ha utilizado la base 2 por lo que la unidad se encuentra en bits. Note que el IG es una *semi-métrica* (Kraskov et al., 2003) que cumple con los axiomas

1. $IG(X;Y) \geq 0$ (no negatividad).
2. $IG(X;Y) = 0 \iff X = Y$ (son independientes).
3. $IG(X;Y) = IG(Y;X)$ (simetría).

Incertidumbre Simétrica. El valor del IG puede ser normalizado utilizando ambas entropías, originando la medida de Incertidumbre Simétrica (SU)(Fayyad and Irani, 1993).

Definition 11. *La incertidumbre simétrica de dos variables aleatorias* X, Y *se define como*

$$SU(X,Y) := 2\left[\frac{IG(X|Y)}{H(X) + H(Y)}\right]. \tag{A.4}$$

Observe que, (a) si X e Y son independientes entonces $SU(X,Y) = 0$; y (b) si X e Y están completamente correlacionados entonces $IG(X|Y) = H(X) = H(Y)$ por tanto $SU(X,Y) = 1$. Como podemos apreciar, el SU restringe sus valores al rango entre 0 y 1, es decir, $SU \in [0,1]$.

Correlación Total. De manera a generalizar la ganancia de información, se introduce la Correlación Total o Multi-información (McGill, 1954; Watanabe, 1960). Esto permite establecer el nivel de correlación de n variables aleatorias que conforman un conjunto.

Definition 12. *Dado un conjunto de n variables aleatorias* $\{X_1, \ldots, X_n\}$*, la correlación total se define como*

$$C(X_{1:n}) := \sum_{i=1}^{n} H(X_i) - H(X_{1:n}), \tag{A.5}$$

donde

$$H(X_{1:n}) := H(X_1, \ldots, X_n) := -\sum_{x_1} \cdots \sum_{x_n} P(x_1, \ldots, x_n) \log_2[P(x_1, \ldots, x_n)] \tag{A.6}$$

es la entropía conjunta de las variables aleatorias $X_1, \ldots, X_n$*.*

A.2.1.3 Incertidumbre Simétrica Multivariada

La incertidumbre simétrica multivariada (MSU) es propuesta en (Sosa-Cabrera et al., 2019) como una generalización del SU basada en la correlación total, con el objeto de cuantificar la redundancia o dependencia existente entre dos o más variables.

Definition 13. *Dado un conjunto de n variables aleatorias* $\{X_1, \ldots, X_n\}$*, la incertidumbre simétrica multivariada se define como*

$$MSU(X_{1:n}) := \frac{n}{n-1} \left[\frac{C(X_{1:n})}{\sum_{i=1}^{n} H(X_i)} \right]. \tag{A.7}$$

Al igual que en el SU, el rango de valores del MSU también se encuentra entre $[0,1]$.

Cabe destacar que a diferencia de la desviación estándar y otras medidas que están orientadas a datos numéricos, la SU y la MSU pueden ser aplicadas a números discretos y a variables aleatorias categóricas. Esta propiedad es conveniente para el mundo multivariado de atributos de diferentes tipos que aparecen frecuentemente en el mismo conjunto de datos.

Además, dado que SU y MSU dependen únicamente de probabilidades, las mismas son invariantes ante traslaciones y cambios de escala aplicados a cualquier X_i, siempre y cuando la función de probabilidad permanezca igual.

A.2.1.4 Divergencia de Kullback-Leibler

La divergencia de Kullback-Leibler (D_{KL} por sus siglas en inglés) es una medida basada en la teoría de la información ampliamente utilizada para calcular la diferencia entre dos distribuciones de probabilidad P y Q que esta dado por

$$D_{KL}(P,Q) := \sum_i P(i) \log\left(\frac{P(i)}{Q(i)}\right) + \sum_i Q(i) \log\left(\frac{Q(i)}{P(i)}\right). \tag{A.8}$$

A.2.1.5 Estrategia de Separabilidad de Clases

En base a la divergencia de Kullback-Leibler, para medir la relevancia y la redundancia de un atributo F con relación a la etiqueta de clase $c_i (c_i \in C)$ dado el subconjunto de atributos seleccionados S, se tiene la denominada estrategia de separabilidad de clases (CS, por sus siglas en inglés) (Zhang et al., 2013) definida mediante

$$Diff_S(F, c_i) := \sum_{f \in S} P_{C_i}(f_S) \cdot D_{KL}(P_{C_i}(F|f_S)||P(F|f_S)). \tag{A.9}$$

A.2.2 Consistencia de un Conjunto de Datos

Las medidas de consistencia, en este sentido, evalúan la irrelevancia colectiva de un conjunto de atributos en función a las etiquetas de clase, es decir, los valores posibles de la variable explicada.

Definition 14. *Un conjunto de atributos de un conjunto de datos es coherente, si, y sólo si, determina unívocamente las clase de cada caso, es decir, dos instancias o ejemplos del conjunto de datos que son idénticos con respecto a los valores de los atributos tienen la misma etiqueta de clase o caso.*

Por lo tanto, una función de medida de consistencia devuelve el valor cero, si, y sólo si, su entrada es un conjunto de atributos consistente (Shin et al., 2017). Un ejemplo importante de la medida de consistencia es la medida de consistencia binaria, definida como sigue:

$$Bn(f_1, \ldots, f_n) := \begin{cases} 0, & \text{si } \{f_1, \ldots, f_n\} \text{ es consistente;} \\ 1, & \text{en otro caso.} \end{cases} \tag{A.10}$$

A.2.3 Selección de Atributos

De ahora en adelante, llamaremos atributo a una variable aleatoria discreta o categórica presente en el conjunto de datos considerado.

La selección de atributos es la tarea de obtener un subconjunto de atributos del menor tamaño posible a partir de los originales presentes en un conjunto de datos dado y que proporcionen la mayor parte de la información útil. Esto es, sin que se vea afectada la predictibilidad de la clase.

Para ello se dejan de lado los atributos detectados como *irrelevantes* como también los que son *redundantes*, efectuando así una reducción de dimensionalidad que resulta en un subconjunto más simple y más adecuado para la predicción deseada.

Además de la simplificación del modelo, la reducción de dimensionalidad lograda a través de una apropiada selección de atributos, disminuye el riesgo de sobreajuste (*overfitting*) del modelo. Un modelo sobreajustado es aquel que ha exagerado su adecuación a los casos de aprendizaje, con el consecuente deterioro en la precisión de sus predicciones para casos nuevos.

Los métodos de selección de características pueden ser clasificados en filtros (*filter*), envolventes (*wrapper*) y embebidos (*embedded*). En los métodos tipo envolvente la evaluación del subconjunto de características se realiza por medio del propio algoritmo de aprendizaje, el cual funciona en este respecto como una especie de caja negra. Esta estrategia posee una alta precisión en cuanto a la calidad de los conjuntos de características, sin embargo, son costosos en términos de recursos computacionales como así también presentan un alto riesgo de sobre-ajuste. En los métodos tipo filtro, la evaluación del subconjunto de características se lleva a cabo por medio de la valoración de las propiedades intrínsecas del dato, tales como la distancia, la consistencia, la entropía y la correlación. Esta estrategia no considera ninguna interacción con el algoritmo de aprendizaje por lo que son mucho más eficientes en recursos computacionales que los métodos tipo envolvente, pero en contrapartida los resultados de la clasificación podrían ser peores. En los métodos tipo embebido la selección de características está incluida en el mismo como una parte no separable, es decir, se realiza la selección de características durante la inducción del clasificador. Los tipos de medidas utilizados en la evaluación de las características como parte del proceso de la selección de los mismos, se distinguen en distancia, medida de información, dependencia estadística, consistencia y error del clasificador. La medida de distancia también es conocida como medida de discriminación o de similitud y la más extendida es la distancia euclídea. El conjunto de atributos a elegir será aquel en el que la separación entre dos regiones sea máxima y la separación entre los casos de la misma clase sea mínima. Los métodos

basados en información utilizan medidas derivadas de la teoría de la información. La información contenida en los atributos es tratada como magnitud física y dicha información se caracteriza mediante la entropía. Los métodos basados en dependencia (o correlación) se basan en el estudio de la relación estadística existente entre los atributos y la clase con el fin de predecir el valor de una en función del valor de la otra. La métodos basados en la medida de consistencia pretende encontrar el subconjunto mínimo de variables con las que es posible construir una hipótesis consistente con el conjunto de entrenamiento. Es decir, los valores de los atributos de dos casos deben ser distintos en caso de que pertenezcan a clases distintas. Finalmente, los métodos basados en la tasa de error emplean el error del clasificador inducido como medida de calidad donde el subconjunto a encontrar es aquel que tenga menor tasa de error en el aprendizaje.

Relevancia de atributos. Sea A el conjunto de todos los atributos, $A_i \in A$ uno de ellos y $S_i = A - \{A_i\}$, como el complemento de A_i respecto de A. Sea C la clase cuya información se desea predecir. A continuación se dará una clasificación de los atributos conforme su relevancia o irrelevancia.

Definition 15 (Relevancia fuerte)**.** *El atributo A_i tiene una relevancia fuerte si y solamente si*

$$P(C|A_i, S_i) \neq P(C|S_i), \tag{A.11}$$

donde $P(C|A_i, S_i)$ es la probabilidad de suponer el valor de C conociendo previamente los valores de A_i y S_i.

Un atributo con relevancia fuerte es aquel que contiene información única acerca de la clase, es decir, no puede descartarse del conjunto A sin alterar la predictibilidad de C.

Definition 16 (Relevancia débil)**.** *El atributo A_i tiene una relevancia débil, si y solamente si,*

$$P(C|A_i, S_i) = P(C|S_i) \tag{A.12}$$

y existe $S'_i \subset S_i$ tal que

$$P(C|F_i, S'_i) \neq P(C|S'_i). \tag{A.13}$$

Un atributo con relevancia débil no es relevante para A, es decir, no cambia la información sobre C si se la descarta de A. Sin embargo, sí es relevante para un subconjunto de A.

Definition 17 (Irrelevancia)**.** *A_i se considera irrelevante, si y solamente si,*

$$P(C|A_i, S_i') = P(C|S_i') \quad \forall S_i' \subseteq S_i. \tag{A.14}$$

Un atributo irrelevante no aporta información alguna sobre C, así debe descartarse para la solución final.

Redundancia de atributos. Se dice que un atributo es redundante si su información está completamente contenida en uno o más atributos.

Definition 18 (Manta de Markov)**.** *Dado un atributo A_i, sea $M_i \subset A - \{A_i\}$. Se dice que M_i es una manta de Markov para A_i si y solamente si*

$$P(A - M_i - \{A_i\} \,|A, M_i) = P(A - M_i - \{A_i\} \,|M_i). \tag{A.15}$$

Una manta de Markov de un atributo es en esencia un conjunto de atributos que contiene toda la información del atributo. Puede verse intuitivamente como una manta que envuelve, oculta al atributo, pues ya aporta toda la información que podría aportar el atributo de los demás elementos de A.

Definition 19 (Redundancia)**.** *Sea $G \subseteq A$ el conjunto actual de atributos. Un atributo se considera redundante (y debe ser descartado de G), si y solamente si, es débilmente relevante y tiene una manta de Markov dentro de G.*

De este modo, se considera un atributo redundante si su información está completamente contenida en un subconjunto de A que no lo incluye.

Un método de selección de atributos del tipo filtro intenta seleccionar el subconjunto de atributos de tamaño mínimo según un bucle de generación de subconjuntos (por estrategia de búsqueda) y su evaluación (por medida) hasta que se satisface algún criterio de parada. A partir de estos pasos básicos, en el Algoritmo 1 se representa un algoritmo abstracto para la selección de atributos que muestra el comportamiento de cualquier método de filtrado de forma unificada.

A.3 Estado del Arte

El objetivo de este trabajo es estudiar el impacto de las interacciones de orden superior en los métodos del tipo filtro para la selección de atributos que utilizan principalmente medidas basadas en la teoría de la información. Por tanto, nos centramos en revisar los métodos en la literatura que adoptan este enfoque.

Algorithm 2: Una generalización del método de filtrado

Entrada: Conjunto de atributos candidatos F, un subconjunto de atributos de partida para la búsqueda S_0 y un criterio de parada δ.

Salida: subconjunto de atributos más informativos S_{mejor}.

```
S_mejor ← S_0  // inicializar S_mejor.
γ_mejor ← evaluar(S_0, F, M)  // evaluar S_0 mediante la medida M.
repetir
    S ← estrategia búsqueda(F, S_mejor)  // generar el siguiente subconjunto.
    γ ← evaluar(S, F, M)  // evaluar S mediante la medida M.
    si γ es mejor que γ_best entonces
        γ_mejor ← γ  // actualizar γ_mejor.
        S_mejor ← S  // actualizar S_mejor.
    fin
hasta que δ se cumple
retorna S_mejor
```

En este sentido, cabe destacar que pocos estudios han examinado la información sobre las interacciones, especialmente las de alto nivel. Dado que es difícil medir directamente la interacción y que las interacciones candidatas crecen exponencialmente con el número de atributos (Tang et al., 2018). Así, en lugar de calcular directamente los términos de interacción de quinto orden, que son costosos computacionalmente, *FJMI* (Tang et al., 2019) tuvo en cuenta las interacciones de segundo a quinto orden entre los atributos y la clase para capturar las interacciones. El enfoque se basa en el hecho de que la información mutua conjunta de cinco dimensiones puede descomponerse en una suma de interacciones de segundo orden, que es más fácil de calcular (Sosa-Cabrera et al., 2023).

En (Vinh et al., 2016) se propone un método de selección de atributos basado en MI de mayor dimensión denominado *RelaxMRMR*. Para capturar las interacciones de atributos de orden superior, los autores identificaron los supuestos que se pueden relajar para descomponer el criterio de información mutua conjunta completa en cantidades MI de menor dimensión (Sosa-Cabrera et al., 2023).

Para tratar explícitamente la interacción entre atributos, en (Zeng et al., 2015a) se propone un método de filtrado basado en la complementariedad denominado *IWFS*. El enfoque se basa en factores de peso de interacción, una variación de la interacción de tercer orden que puede medir la redundancia y la complementariedad entre los atributos (Sosa-Cabrera et al., 2023).

El método *INTERACT* (Zhao and Liu, 2009) encuentra atributos que interactúan basándose en una métrica de ordenación de atributos que utiliza la consistencia de los datos. A diferencia de una evaluación basada en la información mutua, la medida de

inconsistencia es monótona, lo que permite según los autores una búsqueda eficiente para explorar las interacciones entre los atributos (Sosa-Cabrera et al., 2023).

En (Pawluk et al., 2019b) se propone un método de selección de atributos denominado *IIFS* que considera las interacciones de tercer y cuarto orden. Basándose en la información de interacción, demuestran algunas propiedades teóricas del novedoso criterio y la posibilidad de que pueda extenderse al caso de términos de orden incluso superior (Sosa-Cabrera et al., 2023).

Para retener aquellos atributos con la mayor complementariedad con el subconjunto de atributos previamente seleccionado en el progreso del método, en (Li et al., 2020b) se propone un nuevo algoritmo denominado *FS-RRC* que calcula la puntuación de complementariedad de dos atributos y la clase, es decir, una interacción de tercer orden (Sosa-Cabrera et al., 2023).

En (Singha and Shenoy, 2018) se propone un método adaptativo denominado *SAFE* que utiliza una función de costo adaptativo de tercer orden que utiliza la relación redundancia-complementariedad para actualizar automáticamente la regla de compromiso entre relevancia, redundancia y complementariedad. Este enfoque utiliza la estrategia de búsqueda *el primero el mejor*, que ofrece según los autores la mejor solución de compromiso (Sosa-Cabrera et al., 2023).

A.4 Método Propuesto

A diferencia de las técnicas actuales basados en intercooperación de atributos, nuestro método propuesto PART_FS, se basa en la partición del espacio de búsqueda en subespacios y en la aplicación de medidas tanto basado en la información como en consistencia para la detección de dependencias multivariadas no-lineales.

Cabe destacar que el completo detalle del nuevo criterio propuesto aparecerá en una publicación posterior del autor principal[1].

A.5 Resultados Numéricos

El rendimiento del método propuesto PART_FS, se comparó con los resultados de otros 5 métodos métodos: FJMI, IIFS, SAFE, FS_RRC y RELAX_MRMR considerados de tercera generación (Sosa-Cabrera et al., 2023).

Los experimentos fueron realizados sobre 3 escenarios con datos sintéticos y con un total de 20 conjuntos de datos reales del repositorio *UCI* (Markelle Kelly, 2023)

[1] `https://scholar.google.com/citations?user=Wrjw2XYAAAAJ`

de distintos ámbitos: ciencias de la vida, ciencias físicas, ingeniería, negocios, ciencias sociales y otros.

Las características de estos conjuntos de datos pueden ser binarias, discretas, categóricas o continuas. Las características continuas se discretizaron en 10 intervalos iguales utilizando el método de discretización de igual rango. La discretización se realiza como un paso de preprocesamiento para todos los datos antes de la etapa de selección de atributos.

Se utilizaron 7 clasificadores para evaluar la calidad de los subconjuntos seleccionados a saber: *Naive Bayes* (NB), *Support Vector Machine* (SVM), *k-Nearest Neighborhood* (kNN), *Decision Tree based Classification* (J48), *Random Forest* (RF), *Bayes Net* con *K2* para la búsqueda de las estructuras de red (BN_K2) donde 5 es el máximo número de ancestros y *Part Rules Based Classifier* (PRBC). La implementación utilizada de todos los clasificadores se encuentran disponibles en la herramienta WEKA (Eibe et al., 2016).

La precisión media de la clasificación se utiliza como medida de la calidad de los atributos seleccionados. La validación cruzada de 10 iteraciones se emplea al procesar la selección y validación de atributos; por lo tanto, cada muestra de datos se utiliza una vez para la validación.

Los resultados, como se muestra en la figura A.1, indican que los conjuntos de atributos seleccionados por el método propuesto PART_FS permiten una mayor precisión en la tarea de clasificación.

Cabe mencionar finalmente, que el completo detalle de todos los resultados obtenidos aparecerá en una publicación posterior del autor principal[2].

A.6 Conclusiones y Trabajo Futuro

La llegada del Big Data y especialmente el advenimiento de conjuntos de datos con alta dimensionalidad, ha traído una importante necesidad de identificar los atributos relevantes a partir de los datos. En este escenario, la importancia de la selección de atributos está fuera de toda duda y en ese cometido se han desarrollado diferentes métodos, empero a día de hoy, los investigadores no se ponen de acuerdo sobre cuál es el mejor método para un entorno determinado (Bolón-Canedo and Alonso-Betanzos, 2018).

En este trabajo, primero, introducimos la medida de *Incertidumbre Simétrica Multivariada* (*MSU*, por sus siglas en inglés), como una extensión de la *Incertidumbre*

[2] `https://scholar.google.com/citations?user=Wrjw2XYAAAAJ`

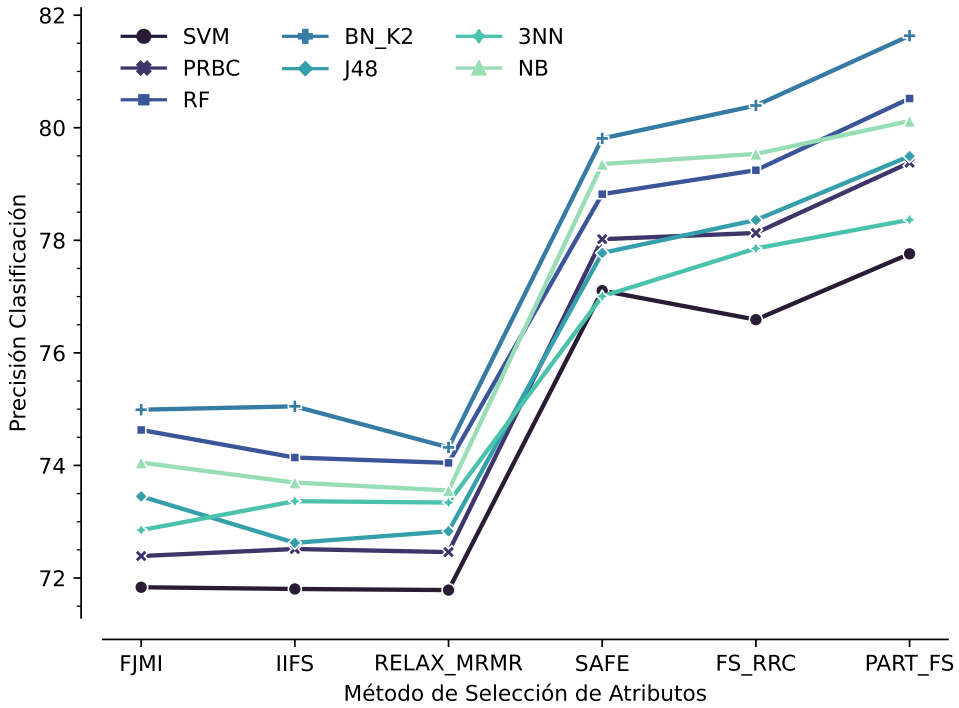


Figure A.1: Precisión de clasificación sobre los atributos seleccionados por PART_FS en comparación a los demás 5 métodos del estado del arte.

Simétrica (*SU*, por sus siglas en inglés) al caso multivariado. Para evaluar la propuesta referida, se han realizado varios experimentos con conjuntos de datos sintéticos y reales. Los resultados han confirmado que el *MSU* es una medida de correlación multivariada confiable para variables nominales, con propiedades prometedoras, capaz de detectar dependencias o interacciones lineales y no-lineales. Para una revisión profunda de este tópico de la tesis, véase la publicación[3] (Sosa-Cabrera et al., 2019).

En adición, proporcionamos un estudio sistemático del estado del arte sobre la asistencia y explotación de la intercooperación de atributos en el proceso de selección de atributos. En este sentido, hemos examinado un total de 27 métodos filtro de selección de atributos que adoptan este enfoque identificado, cubriendo lagunas importantes en el campo de los métodos de última generación, un tema que hasta ahora no había recibido mucha consideración en la literatura. Para acceder al examen bibliográfico exhaustivo, véase la publicación[4] (Sosa-Cabrera et al., 2023).

Y, finalmente, se propone un enfoque novedoso de selección de atributos basado en la partición del espacio de búsqueda de atributos y la intercooperación de atributos denominado PART_FS. PART_FS es un marco particularmente versátil para datos de

[3] `https://doi.org/10.1016/j.ins.2019.04.046`
[4] `https://doi.org/10.1007/s41060-023-00439-z`

alta dimensión de naturaleza compleja. En este sentido, comparamos el rendimiento de PART_FS en escenarios simulados y conjuntos de datos reales con varios métodos recientes de selección de atributos en combinaciones con diferentes clasificadores. Los resultados muestran que el método propuesto basado en la partición y la intercooperación supera a los métodos de comparación y sobresale en una variedad de problemas con diferentes características. Un completo abordaje sobre este tópico aparecerá en una publicación posterior del autor principal[5].

Sin embargo, la selección de atributos sigue siendo y seguirá siendo un campo activo que se rejuvenece incesantemente para responder a nuevos desafíos (Liu et al., 2010). Por ejemplo, dado que la precisión de MSU depende de muestras que sean totalmente representativas, un principal inconveniente de esto (tamaño de muestra basado en representatividad total) consiste en el hecho de que el tamaño de la muestra aumenta con la cardinalidad multivariada. Esto implica tamaños de muestra más grandes para lograr una precisión prescrita. Por otra parte, el rendimiento de PART_FS podría mejorarse aún más examinando cuidadosamente las atributos de los conjuntos de datos reales, modificando el criterio de partición y optimizando los parámetros del modelo en consecuencia. El trabajo futuro se centrará en estos puntos.

[5] `https://scholar.google.com/citations?user=Wrjw2XYAAAAJ`