Bayesian Machine Learning Techniques for revealing complex interactions among genetic and clinical factors in association with extra-intestinal Manifestations in IBD patients.

E. Menti, MS[1], C. Lanera, MS[1], G. Lorenzoni, MS[1], Daniela F. Giachino, MD[2], Mario De Marchi, MD[2], Dario Gregori, PhD MA[1], Paola Berchialla, PhD MA[3] and Piedmont Study Group on the Genetics of IBD *

[1]Unit of Biostatistics, Epidemiology and Public Health, University of Padova, Italy; [2]Medical Genetics Unit, Department of Clinical and Biological Sciences, University of Torino, Italy; [3]Medical Statistics Unit, Department of Clinical and Biological Sciences, University of Torino, Italy

*Marco Astegiano, Nicoletta Sapone, Elena Terzi -Gastrohepatology D1 Unit, San Giovanni Battista Hospital, Torino (Director prof Mario Rizzetto);
Angela Sambataro, Paola Salacone, Ezio Gaia - Gastroenterology Unit, San Luigi Hospital, Orbassano;
Rodolfo Rocca, Alessandro Lavagna, Lucia Crocellà, Annalisa Vernetto, Marco Daperno, Angelo Pera - Gastroenterology Unit, Ordine Mauriziano Hospital, Torino;
Silvia Regazzoni, Marco Bardessono - Medical Genetics Unit, Department of Clinical and Biological Sciences, University of Torino

Abstract

The objective of the study is to assess the predictive performance of three different techniques as classifiers for extra-intestinal manifestations in 152 patients with Crohn's disease. Naïve Bayes, Bayesian Additive Regression Trees and Bayesian Networks implemented using a Greedy Thick Thinning algorithm for learning dependencies among variables and EM algorithm for learning conditional probabilities associated to each variable are taken into account. Three sets of variables were considered: (i) disease characteristics: presentation, behavior and location (ii) risk factors: age, gender, smoke and familiarity and (iii) genetic polymorphisms of the NOD2, CD14, TNFA, IL12B, and IL1RN genes, whose involvement in Crohn's disease is known or suspected. Extra-intestinal manifestations occurred in 75 patients. Bayesian Networks achieved accuracy of 82% when considering only clinical factors and 89% when considering also genetic information, outperforming the other techniques. CD14 has a small predicting capability. Adding TNFA, IL12B to the 3020insC NOD2 variant improved the accuracy.

Keywords: *Clinical Decision Support, Clinical research informatics, Data mining and statistical data analysis.*

Introduction

The extensive clinical heterogeneity of Inflammatory Bowel Disease (IBD), and in particular of Crohn's disease (CD), has stimulated several efforts to classify patients according to recognized criteria, from the international meetings in Rome (1991) and Vienna (1998) to the recent revision in Montreal [1]. The presence of extra-intestinal manifestation (EIM), among others, has important consequences for the clinical management of CD patients and relevant effects on the overall burden of the disease, the quality of life and the allocation of health resources [2]. Attempts have been made to weight the risk of EIM according to the patients' conditions both at onset and during disease progression, and among the potential risk factors for onset of EIM the role of several genetic predisposing/modifier factors have been recently reviewed [3], even if their clinical usefulness is at present unclear [4].

This increase of information, usually in conjunction with the limited size of the analyzed samples, is posing several threats to the statistical procedures used for EIM risk stratification. Classical and most used tools, such as logistic

models, are known to have limits in such situations. In addition, the effect of some risk factors on the risk of EIM is known to be non-linear and to interact with other covariates [5, 6].

In a previous paper from Giachino et al. [7], six common statistical models (logistic regression model, generalized additive models, linear and quadratic discriminant analysis, artificial neural networks (ANN) and projection pursuit regression (PPR) were implemented to predict EIM using genetic data in addition of clinical factors, showing the impact that genetics, when appropriately modeled, can have in predicting EIM.

The aim of this paper is to further develop on that pathway, approaching the problem of "predicting" EIM by implementing three different Bayesian classifiers and by assessing its predictive capability in comparison with these, previous and current, results.

Several approaches have been proposed in the Bayesian framework to deal with classification (or prediction) in a clinical setting. Among them, of a heuristically increasing complexity, major groups include naïve Bayes, Bayesian Additive Regression Trees (BART) and Bayesian networks.

Naïve Bayes (NB) classifier applies Bayes' theorem by assuming that the features are independent given class, regardless of any possible correlation between them. Studied intensively from 1950s, it has been widely adopted in automatic medical diagnosis, with convincing performances, often outperforming other sophisticated techniques, despite its sometime unrealistic independence assumption [8].

Bayesian Additive Regression Trees (BART), having been developed at the beginning for regression problems, is a nonparametric statistical approach making use of a sum-of-trees model and regularization prior on the parameters in order to approximate an unknown function. It has been extended by Chipman et al. [9] to the probit model setup to handle binary classification tasks.

Bayesian Networks (BN) have been introduced in the 1980s as a probabilistic expert system for representing and reasoning models of problems involving uncertainty. Since the beginning of the 1990s, they have been used for developing medical applications [10, 11]. Their success in this field is due to the fact they possess the quality of being both a statistical and an Artificial Intelligent knowledge-representation tool. Furthermore, they allow for structuring domain knowledge by investigating causal relationships among domain variables [12]. In many cases, Bayesian Networks have been proven to outperform other statistical methodology in classification tasks [13].

Materials and Methods

The present dataset derives from the larger series of CD and Ulcerative Colitis patients enrolled in our ongoing observational study of IBD genetics in collaboration with three gastroenterology Units in Torino, Italy. An association analysis of the three common NOD2 variants has been reported. Genomic DNA was extracted using a commercial kit (Promega). The nomenclature of the analysed polymorphisms is reported in Table 1, together with references to typing technique and relevant literature. Of the two polymorphisms in the 5' region of the TNFA gene we here consider only the genotype at -308, since all analysed samples were homozygous for the common G allele at the -238 SNP.

Table 1. Analyzed SNPs.

Gene	Polymorphism	Analysis	dbSNP ID[32]	Pr F	Pr R	Restriction enzyme		Ref.
NOD2	R702W	PCR-RFLP	rs2066844	5'-AGGTCA-GCCTGATG-ACATTTC-3'	5'-CGGGAT-GGAGTGG-AAGT-3'	Msp I	A:329+66+54 bp T:329+120bp	Giachino et al 2004[7]
	G908R	PCR-RFLP	rs2066845	5'-CACTGA-CACTGTCT-GTTGACTC-3'	5'-AAGACC-TTCAGAAC-TGGCCCC-3'	HhaI	G: 202bp C:155+47bp	Giachino et al 2004[7]
	INSC3020	PCR-RFLP	rs2066847	5'-CTGGCT-AACTCCTG-CAGT-3'	5'-ACTGAG-GTTCGGA-GAGCT-3'	NlaIV	insC: 142+37+38 bp wt: 180+37bp	Giachino et al 2004[7]
CD14	-159C>T	PCR-RFLP	rs2569190	5'-GTGCCA-ACAGATGA-GGTTCAC-3'	5'-GCCTCT-GACAGTTT-ATGTAATC-3'	AvaII	T: 353 + 144bp C: 479bp	Klein 2002[33]

	continued							
TNFA	-308G>A	ARMS-RFLP	rs1800629	5'-AGGCAA-TAGGTT-TTGAGGGG-CAT-3'		NcoI	A: 117bp G: 97 + 20bp	Vinasco et al. 1997[8]
	-238G>A	ARMS-RFLP	rs361525		5'-ACATCC-CCATCCTC-CCAGATC-3'	BglII	G: 117bp A: 97 + 20bp	Vinasco et al. 1997[8]
IL12B	Ex8 +159A>C	PCR-RFLP	rs3212227	5'-TTTGGA-GGAAAAGT-GGAAGA-3'	5'-AACATT-CCATACAT-CCTGGC-3'	TaqI	C: 161+139bp A: 300bp	D'Alfonso (personal communication)
IL1RN	86bp VNTR	Elecrophoresis	AJ289235	5'-CTCAGCAA-CACTCCTAT-3'	5'-TCCTGGTC-TGCAGGTAA-3'	-	-	Mansfield et al. 1994[34] Vamvakopoulos 2002[35]

Patients

We decided to use for this work the same data set as in the previous analysis [7] in order to allow a direct comparison of the various statistical approaches. Detailed clinical and familiar information were acquired from each patient and encoded according to the Vienna classification that was in use at the time of enrolment. Extra-intestinal manifestations were defined as the occurrence of rheumatologic, dermatological, ocular, liver and biliary manifestations and amyloidosis. Patients form a retrospective cohort belonging to the Italian Population. They gave a written consent to the study, which was performed under permission of the Hospital Ethical Committee.

Naïve Bayes

A Naïve-Bayes classifier [14] is a simple BN that has the outcome variable as the parent node of all other nodes and no other connections between variables.

Over the BN's they are easy to construct, since the structure is given a priori and thus no structure learning procedure is needed. They require the assumption that all the features are independent of each other. Despite this strong assumption, Naive-Bayes have proven to outperform many classifiers especially where the features are not strongly correlated [15].

The naïve classifier combines a probability model with a decision rule: it computes the conditional a *posterior* probabilities of a categorical class variable given independent predictor variables using the Bayes' theorem. The metric predictors are supposed to be distributed as a Gaussian. This technique is the simplest class of Bayesian Networks where all of the features are class-conditionally independent. Its simplicity makes it easy to use and it allows to get a good result especially in case of small databases [8]. Moreover, naïve classifiers can be extremely fast in comparison to more sophisticated methods of data mining.

Naïve Bayes classifier's implementation during this studio takes advantage of the "e1071" R package [16].

Bayesian Additive Regression Trees

A Bayesian Additive Regression Trees is a nonparametric Bayesian approach to estimation which uses dimensionally adaptive random basis elements, the regression trees, to approximate an unknown function $f(x) = E(Y|x)$, by imposing accurately the regularization prior. By weakening the single effects BART ends up with a sum of trees each of which explains a small and different portion of the function f [9]. Obviously, respect to single trees, models composed of sums of trees have a greater ability to describe into details f, capturing interaction and non-linearity. Hence the information regarding f is partitioned into different trees, each contributing to the overall fit.

This technique was primarily designed to predict quantitative (continuous) outcomes from observations via regression, but an algorithm that extends BART for binary classification, written in the statistical R package "BayesTree", is provided online by the original authors of BART [9].

In general, as anticipated, BART model consists of two parts: a sum-of-trees model and a regularization prior of the model's parameters that keeps the individual tree effects small. Mathematically BART probit extension model for binary classification (coded with outcomes "0" and "1") can be expressed as:

$$P[\boldsymbol{Y} = 1|\boldsymbol{X}] = \phi[\mathcal{T}_1^M(\boldsymbol{X}) + \mathcal{T}_2^M(\boldsymbol{X}) + \cdots + \mathcal{T}_m^M(\boldsymbol{X})]$$

where $\phi[\bullet]$ denotes the cumulative density function of the standard distribution, $\mathcal{T}$ denotes a binary tree made of a set of node decision rules and a set of terminal nodes, M a set of parameter values associated with each of the terminal nodes of $\mathcal{T}$. The number of trees, m, is fixed to 200 by choice as a good trade-off: the more this number increases the more the model is flexible, showing excellent prediction capabilities slowing down the computational time.

Through a Bayesian backfitting MCMC algorithm [17] that iteratively constructs and fits successive residuals, thanks to the data augmentation approach of Albert and Chib (1993) [18], the posterior information is extracted. MCMC chooses between different generated trees the one providing the best sum-of-trees model according to a *posterior* probability. This approach, because of the complex computations, is usually time consuming. However, BART has several appealing features such as the additive characteristic able to catch the variability of the function and the capability to conduct automatic variable selection.

BART approach releases the strong hypothesis of statistical independence of attributes making this technique more akin to real data.

Bayesian Networks

A Bayesian Network is a graphical representation of the joint probability distributions over a set of random variables. It consists of a series of nodes representing variables connected by arrows forming a graph that has no cycles. The arcs specify the independence assumptions that must hold between the random variables.

In general, they may be many arcs going into and out of each node, creating a complex network. The most important restriction is that the arcs must not create cycles within the network; the resulting network is known as directed acyclic graph (DAG) [19]. Each node of the network is associated with a set of probability tables. For those nodes without ingoing arcs, the probability distribution is a prior distribution which requires supplying a set of initial values. Both the structure and the numerical parameters of a BN can be learned entirely from data [20, 21].

There are a great number of algorithms for learning the structure and the parameters of Bayesian networks from data. Many of them are based on a scoring function and a search procedure. The algorithms based on a scoring function try to find a graph that best represents the data, according to a specific criterion. They use a scoring function in combination with a search method to measure the goodness of each explored structure from the space of feasible solutions. During the exploring process, the scoring function is applied to evaluate the fitness of each candidate structure to the data.

In this analysis, a variant of this scoring approach is the Greedy Thick Thinning algorithm [22], which optimizes an existing structure by modifying the structure and scoring the result, was performed. By starting from a fully connected DAG and subsequently removing arcs between nodes based on conditional independences tests [23], the Greedy Thick Thinning algorithm is able to isolate the best scoring network. One of the most usual scoring function is the Bayesian metric [24], which is a measure of how likely it is to observe the data given the network structure, i.e. the best network in terms of the Bayesian metric is that one with the highest probability based on the given data [24].

Given the structure of the network, conditional probability learning is done. Since conditional probabilities to be learned depend not just on the parent variables' values but also on the other linked variable (local structure), usually the assumption each variable is discrete is made. In this way, each local distribution is a collection of multinomial distributions. Given this class of local distributions, probabilities can be efficiently computed when there are no missing data in the sample and assuming local parameter independence, i.e. the probability of each state is independent of the probability of every other state. The learning method performed in this analysis was Expectation-Maximization (EM) algorithm [25].

The EM algorithm performs a number of iterations. For each iteration, the logarithm of the probability of the case data given the current joint probability distribution is computed and the EM-algorithm attempts to maximize this quantity. The starting point of the EM algorithm is the conditional probability tables specified prior to calling the algorithm. As a *priori* distribution, the uniform distribution was assumed for each variable. The EM algorithm

terminates when the relative difference between the log-likelihood for two successive iterations is sufficiently small (less then10^{-4}).

To assess model performance, error rate and predictive value of the Bayesian Network were estimated using a 10-fold cross validation procedure.

Bayesian Networks implementation was carried out using GeNIe 2.0 [26].

Results

Basic characteristics of the sample, stratified by occurrence of EIM, are presented in Table 2. A more detailed description of the dataset is given in Giachino's work [27]. The clinical and genetic characteristics were divided into three groups: (1) characteristics of the disease: age at onset, location, disease behavior and presentation of the disease; (2) known risk factors: sex, smoking behavior and familiarity of the disease; (3) genetic polymorphisms of the NOD2, CD14, IL12B, TNF, IL1RN genes.

Table 2. Data description. Median, I, III quartile, number, percentages as appropriate. N indicates the number of cases with a valid information for the given covariate.

Variable		N	EIM- (77)	EIM+ (75)	Combined (152)
Characteristics of the disease					
PRESENTATION	Medical	147	70 (90.9%)	63 (84%)	133 (87.5%)
	Surgical		7 (9.1%)	7 (9.3%)	14 (9.2%)
BEHAVIOUR	Nonstricturing, nonpenetrating	108	25 (32.5%)	25 (33.3%)	50 (32.9%)
	Stricturing		15 (19.5%)	20 (26.7%)	35 (23%)
	Penetrating		9 (11.7%)	14 (18.7%)	23 (15.1%)
LOCATION	Terminal ileum	109	14 (18.7%)	14 (18.7%)	28 (18.4%)
	Colon		11 (14.3%)	14 (18.7%)	25 (16.4%)
	Ileocolon		21 (27.3%)	27 (36%)	48 (31.6%)
	Upper GI		4 (5.2%)	4 (5.3%)	8 (5.3%)
AGE	<40 yrs	146	53 (68.8%)	51 (68%)	104 (68.4%)
	>40 yrs		21 (27.3%)	21 (28%)	42 (27.6%)
Risk factors					
SEX	Male	152	46 (59.7%)	34 (45.3%)	80 (52.6%)
	Female		31 (40.3%)	41 (54.7%)	72 (47.4%)
SMOKER	No	146	42 (54.5%)	36 (48%)	78 (51.3%)
	Yes		19 (24.7%)	26 (34.7%)	45 (29.6%)
	Ex smoker		11 (14.3%)	12 (16%)	23 (15.1%)
FAMILIARITY	No	139	58 (75.3%)	57 (76%)	115 (75.7%)
	Yes		11 (14.3%)	13 (17.3%)	24 (15.8%)
Polymorphisms					
NOD2: R702W	RR	152	63 (81.8%)	64 (85.3%)	127 (83.6%)
	RW		11 (14.3%)	9 (12%)	20 (13.2%)
	WW		3 (3.9%)	2 (2.7%)	5 (3.3%)
G908R	GG	152	73 (94.8%)	67 (89.3%)	140(92.1%)
	GR		4 (5.2%)	8 (10.7%)	12 (7.9%)
INSC3020	LL	152	71 (92.2%)	65 (86.7%)	136 (89.5%)
	L/insC		5 (6.5%)	8 (10.7%)	13 (8.6%)

continued					
CD14	CC	152	20 (26.0%)	20 (26.7%)	40 (26.3%)
	TC		39 (50.6%)	36 (48%)	75 (49.3%)
	TT		18 (23.4%)	19 (25.3%)	37 (24.3%)
TNFA -308	GG	72	35 (45.5%)	18 (24%)	53 (34.9%)
	GA		9 (11.7%)	4 (5.3%)	13 (8.6%)
	AA		5 (6.5%)	1 (1.3%)	6 (3.9%)
TNFA -238	GG	72	49 (63.6%)	23 (30.7%)	72 (47.4%)
IL12B	AA	72	17 (22.1%)	11 (14.7%)	28 (18.4%)
	AC		24 (31.2%)	10 (13.3%)	34 (22.4%)
	CC		8 (10.4%)	2 (2.7%)	10 (6.6%)
IL1RN	ILRN*1	72	29 (37.7%)	12 (16%)	41 (27%)
	ILRN*1/ILRN*2		15 (19.5%)	7 (9.3%)	22 (14.5%)
	ILRN*2		3 (3.9%)	3 (4%)	6 (3.9%)
	ILRN*1/ILRN*3		1(1.3%)	1 (1.3%)	2 (1.3%)
	ILRN*2/ILRN*3		1 (1.3%)	0	1 (0.7%)

Due to the high number of patients with missing values, two different datasets are considered: the first one (1) in which TNF and IL1RN genes were excluded and a second one (2) containing only patients for whom data on TNF and IL were available (Figure 1).

The Bayesian Networks were depicted in Figure 1. Arrows between nodes denoted probability dependencies among variables (bolder arrows pointed out stronger influences among variables).

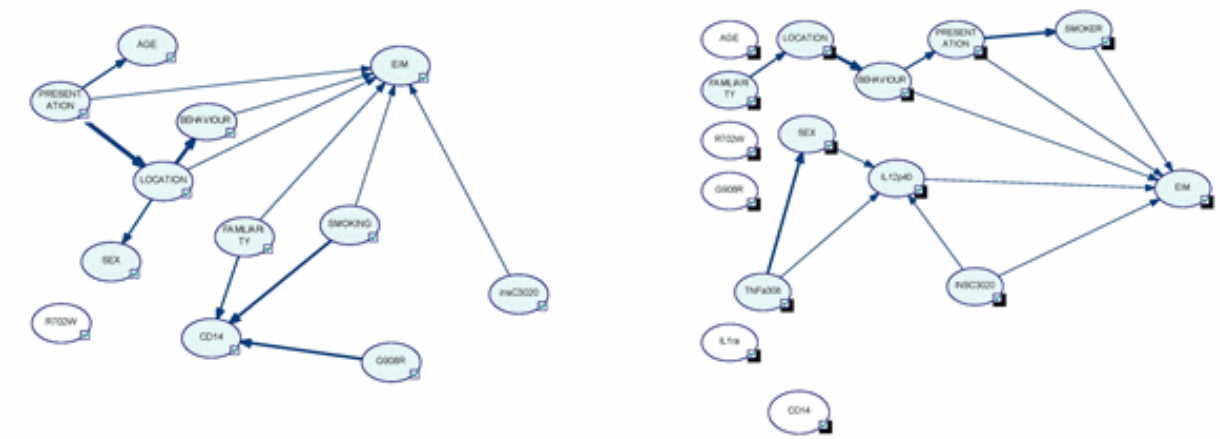

Figure 1. Bayesian Network for characterizing EIM: BN1 (left) and BN2with IL and TNF genes (right).

Sensitivity to finding for BN1 is shown in Table 3. As can be seen from the low value of mutual information and variance of beliefs and from the graph, the role of CD14, is negligible for predicting EIM.

Table 3. Sensitivity of finding analysis for Extra intestinal manifestation in BN1.

Node	Mutual Information	Variance of beliefs
EIM	0.99937	0.2497803
SMOKER	0.00767	0.0026439
LOCATION	0.0047	0.0016265
FAMILIARITY	0.00423	0.0014645
BEHAVIOUR	0.00338	0.0011712
PRESENTATION	0.00325	0.0011018
CD14	0.00095	0.0003287
AGE	0.00033	0.0001152
SEX	0.00008	0.0000286
INSC3020	0.00001	0.0000036
G908R	0	0
R702W	0	0

In Table 4, comparison of accuracy and predictive values of the three statistical approaches implemented, and of their performance's enhancements due to the role of genetic variables, was shown.

Table 4. Accuracy, sensitivity, specificity, positive predictive values (PPV) and negative predictive value (NPV) for the three different techniques in case of considering (2) or not (1) the genetic variables.

	Accuracy	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**AUC**
Naïve Bayes1	0.62 (0.54-0.70)	0.57 (0.45-0.69)	0.66 (0.55-0.77)	0.62 (0.50-0.74)	0.61 (0.50-0.72)	0.64 (0.55-0.73)
Naïve Bayes2	0.79 (0.68-0.88)	0.52 (0.31-0.73	0.92 (0.80-0.98)	0.75 (0.48-0.93)	0.80 (0.68-0.90)	0.78 (0.67-0.90)
BART1	0.64 (0.56-0.71)	0.63 (0.51-0.74)	0.65 (0.53-0.75)	0.64 (0.52-0.74)	0.64 (0.52-0.75)	0.71 (0.62-0.79)
BART2	0.75 (0.63-0.84)	0.26 (0.10-0.48)	0.98 (0.89-1)	0.86 (0.42-1)	0.74 (0.61-0.84)	0.86 (0.78-0.95)
BN1	0.82 (0.76-0.88)	0.84 (0.77-0.91)	0.79 (0.7-0.88)	0.83 (0.69-0.97)	0.80 (0.69-0.91)	0.85 (0.77-0.93)
BN2	0.89 (0.81-0.97)	0.78 (0.71-0.86)	0.94 (0.86-1)	0.86 (0.71-0.95)	0.90 (0.81-0.99)	0.95 (0.86-1)

Discussion

The aim of our study was to compare the performance of three Bayesian classifiers in predicting EIM. Among Bayesian classifiers, in this work we focused on NBs, BNs and BARTs, which are the most popular ones. NBs are often regarded as a benchmark and generally well performing models in spite of their simplicity. On the other hand, BNs can be seen as a more general extension of NB, since they can capture also the interaction between features. Finally, BART represents a full Bayesian classifier.

Our analyses showed that BNs outperformed NB and BART. All classifiers show enhancements when introducing the knowledge about IL and TNF genes, paying the price of a small sensitivity. This comparison is useful to understand the important role of this factor in classification tasks. Taking into account the simplicity and the unrealistic assumption of independence at the basis of NB the results obtained through this classifier are quite comparable to those of BART technique that is more complex and computationally heavier. Typically, the performance of Naïve-Bayes can be further improved by carrying out features selection or by relaxing the independence assumption. From a clinical view a potential explanation could be due to the fact the information about genes is very specific. The involvement in Chron Disease of the genetic polymorphism analyzed is not proved but only suspected. Since this disease has a multifactorial origin, the presence of the polymorphism is not necessarily linked to the presence of the extra-intestinal manifestation but the absence can help in excluding it. This also specifically means that when the information about genes is known the positive predictive values are higher, since better specificity lead to less false positives.

Combining naïve Bayes with features selection is known as selective naïve Bayes [29]. The search strategies for features selection can be carried out following two different approaches: (i) the filter approach and (ii) the wrapper approach [30]. In the filter approach the search strategy is aimed at maximizing the accuracy of the classifier looking only at the discrimination power of the single variables. This is done considering the mutual information function, which is a function independent of the classifier, i.e. the variables already added to the classifier.

Relaxing the independence assumption is basically performed by constructing a tree augmented Naïve-Bayes (TAN), i.e. first learning a structure tree over the set of variables and then adding a link from the response variable to each node, similar to a Naïve-Bayes structure. However, as discussed in [31], NBs often performs well even when the assumption is violated [31].

Finally, NBs rely also on the assumption that continuous variables are Gaussian. The Gaussian assumption means that the conditional probability of each feature given the class is normal with class conditional mean and variance and then it uses maximum likelihood approach to estimate parameters. Since in our data all variables were categorical, the Gaussian assumption was indeed not required.

The present analysis showed that Bayesian Networks were able to provide a further improvement with respect to other statistical model in terms of predictive accuracy. In addition to these features, the graphical nature of BNs allows to display the links between variable. This can facilitate discussion of the model from different backgrounds point of view (clinicians and genetists, for example) and can encourage interdisciplinary. Another important feature of BNs is the ability to learn about the structure and parameters on the basis of observed data. Knowledge of the structure reveals the dependencies of variables and can suggest a direction of causation. However, when using learning algorithm, causal interpretation of dependencies can be a matter of concern and it is more appropriate referring to them as probabilistic relationships.

The assessment of the optimal BN structure is based on the highest probability score for possible candidate structures, given the data provided and eventually penalized for the level of complexity. Different score metrics can be used at this purpose, varying from entropy methods to genetic algorithms. In our analysis we considered a Bayesian metric. The choice of the Bayesian metric as scoring function may lead to have less prediction error compared with BNs suggested by other scoring metrics. However, on the other hand, it may suggest a model that is highly complex and more difficult to interpret with a large number of variables and a large number of links.

A sufficient number of observations is needed to enable a robust estimation of conditional probabilities, even if it has been shown that BNs can yield good prediction accuracy using learning algorithms, even if sample size is small.

As expected for previous studies, our analysis confirms the negligible role of CD14 and that adding the INF-308 and IL12B genotypes information improved the predictive performance of the model. Whereas, adding NOD2, only the variant 3020insC seemed to be associated to predictive accuracy of EIM.

The small sample size and the absence of an independent sample to perform an external validation represent the major limitations of this work, even if the study's conclusions are strengthened internal validation performed with cross-validation procedure, in order to reduce overfitting bias.

Among non-Bayesian classifiers, Projection Pursuit Regression, Artificial Neural Network and Quadratic Discriminant Analysis outperform in AUC the other models. However, BNs resulted in a comparable accuracy with them. Furthermore, BNs has the advantage to provide an interpretable graphical model, which can be easily discussed and accordingly modified on the basis of medical knowledge of the problem.

In this work we did not focus on the impact of the prior distributions over the posterior probabilities. Usually, the Naïve Bayes approach assumes equi-probable classes as a prior or it uses an estimate for the class probability given by the number of samples in the class over the total number of samples. In absence of prior information from other independent studies, we chose the former strategy in order to use data only once, i.e. just in the training/testing step and not also for specifying the a-priori probability over the classes. Of course as the sample size tend to be large, the prior is forgotten and the data play the most important role is taken by the data.
Also the two BNs were learned in a non-Bayesian way using the K2 greedy search algorithm, which has been shown to outperform other algorithms (Comparison of the Bayesian Network Structure Learning Algorithms).
Regarding BART, [32] proposed a method for incorporating informed prior information about the predictors into the BART model by modifying the prior on the splitting rules as well as the corresponding calculations in the Metropolis-Hastings step. In particular, covariates believed to influence the response can be proposed a priori more often as candidates for splitting rules. Also for this classifier, we chose to use uninformative prior, considering to work in a total ignorance situation.

The presence of small datasets is a common situation in medical applications. BN has some interesting implications for clinical practice where the dataset is usually very small, affecting statistical analysis. BN, even if temporal knowledge is not considered explicitly, can be useful to make a prediction about what will happen in future. In fact the clinical available knowledges about the patients before the treatments is started influence the treatments actions and hence the final outcome. The prognostic Bayesian networks importance in health-care is a well-documented topic and the obtained results in terms of accuracy demonstrates their possible employment is automatic medical prediction.

Conclusion

Our study shows that BNs are a feasible and accurate tool for predicting EIM in CD patients. IL and TFN genes influence the classification and bring to a more reliable classification, increasing the accuracy of about 10%.

References

1. Silverberg, M.S., et al., *Toward an integrated clinical, molecular and serological classification of inflammatory bowel disease: Report of a Working Party of the 2005 Montreal World Congress of Gastroenterology.* Can J Gastroenterol, 2005. **19**(Suppl A): p. 5-36.
2. Caprilli, R., et al., *European evidence based consensus on the diagnosis and management of Crohn's disease: special situations.* Gut, 2006. **55**(suppl 1): p. i36-i58.
3. Danese, S., et al., *Extraintestinal manifestations in inflammatory bowel disease.* World Journal of Gastroenterology, 2005. **11**(46): p. 7227.
4. Stange, E., et al., *European evidence based consensus on the diagnosis and management of Crohn's disease: definitions and diagnosis.* Gut, 2006. **55**(suppl 1): p. i1-i15.
5. Cho, J.H., et al., *Identification of novel susceptibility loci for inflammatory bowel disease on chromosomes 1p, 3q, and 4q: evidence for epistasis between 1p and IBD1.* Proceedings of the National Academy of Sciences, 1998. **95**(13): p. 7502-7507.
6. van Heel, D.A., et al., *The IBD6 Crohn's disease locus demonstrates complex interactions with CARD15 and IBD5 disease-associated variants.* Human molecular genetics, 2003. **12**(20): p. 2569-2575.
7. Giachino, D.F., et al., *Modeling the role of genetic factors in characterizing extra-intestinal manifestations in Crohn's disease patients: does this improve outcome predictions?* Current medical research and opinion, 2007. **23**(7): p. 1657-1665.
8. Rish, I. *An empirical study of the naive Bayes classifier*. in *IJCAI 2001 workshop on empirical methods in artificial intelligence*. 2001. IBM New York.
9. Chipman, H.A., E.I. George, and R.E. McCulloch, *BART: Bayesian additive regression trees.* The Annals of Applied Statistics, 2010: p. 266-298.
10. Lucas, P.J., L.C. van der Gaag, and A. Abu-Hanna, *Bayesian networks in biomedicine and health-care.* Artificial intelligence in medicine, 2004. **30**(3): p. 201-214.
11. Mani, S., M. Valtorta, and S. McDermott, *Building Bayesian network models in medicine: The MENTOR experience.* Applied Intelligence, 2005. **22**(2): p. 93-108.
12. Spiegelhalter, D.J., et al., *Bayesian analysis in expert systems.* Statistical Science, 1993: p. 219-247.

13. Foltran, F., et al., *A systems biology approach: new insights into fetal growth restriction using Bayesian Networks.* Journal of biological regulators and homeostatic agents, 2010. **25**(2): p. 269-277.
14. Duda, R.O. and P.E. Hart, *Pattern classification and scene analysis*. Vol. 3. 1973: Wiley New York.
15. Langley, P., W. Iba, and K. Thompson. *An analysis of Bayesian classifiers*. in *Aaai*. 1992.
16. Dimitriadou, E., et al., *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.5-27*. 2011.
17. Hastie, T. and R. Tibshirani, *Bayesian backfitting (with comments and a rejoinder by the authors.* Statistical Science, 2000. **15**(3): p. 196-223.
18. Albert, J.H. and S. Chib, *Bayesian analysis of binary and polychotomous response data.* Journal of the American statistical Association, 1993. **88**(422): p. 669-679.
19. Jensen, F.V., *Bayesian Networks and Decision Graphs*. 2001: Springer.
20. Pearl, J., *Causality: Models, Reasoning and Inference*. 2000, Cambridge: Cambridge University Press.
21. Cooper, G.F. and E. Herskovits, *A Bayesian method for the induction of probabilistic networks from data.* Machine learning, 1992. **9**(4): p. 309-347.
22. Dash, D. and M.J. Druzdzel. *Robust independence testing for constraint-based learning of causal structure*. in *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*. 2002. Morgan Kaufmann Publishers Inc.
23. Spirtes, P., C. Glamour, and R. Scheines, *Causation, Prediction, and Search*. 1993, New York: Springer.
24. Heckerman, D., D. Geiger, and D.M. Chickering, *Learning Bayesian networks: The combination of knowledge and statistical data.* Machine learning, 1995. **20**(3): p. 197-243.
25. Lauritzen, S.L., *The EM algorithm for graphical association models with missing data.* Computational Statistics & Data Analysis, 1995. **19**(2): p. 191-201.
26. Decision Systems Laboratory, *GeNIe 2.0,* http://www.sis.pitt.edu/~genie/. 2006.
27. Giachino, D., et al., *Analysis of the CARD15 variants R702W, G908R and L1007fs in Italian IBD patients.* Eur J Hum Genet, 2004. **12**(3): p. 206-12.
28. Giachino, D.F., et al., *Modeling the role of genetic factors in characterizing extra-intestinal manifestations in Crohn's disease patients: does this improve outcome predictions?* Curr Med Res Opin, 2007. **23**(7): p. 1657-65.
29. Kononenko, I. *Semi-naive Bayesian classifier*. in *Machine Learning—EWSL-91*. 1991. Springer.
30. Blanco, R., I. Inza, and P. Larranaga, *Learning Bayesian networks in the space of structures by estimation of distribution algorithms.* International journal of intelligent systems, 2003. **18**(2): p. 205-220.
31. Domingos, P. and M. Pazzani, *On the Optimality of the Simple Bayesian Classifier under Zero-One Loss.* Mach. Learn., 1997. **29**(2-3): p. 103-130.
32. Bleich, J. and A. Kapelner, *Bayesian Additive Regression Trees With Parametric Models of Heteroskedasticity.* arXiv preprint arXiv:1402.5397, 2014.