The Multilabel Naive Credal Classifier

Alessandro Antonucci and **Giorgio Corani**
IDSIA SUPSI/USI
Lugano (Switzerland)
{alessandro,giorgio}@idsia.ch

Abstract

We present a credal classifier for multilabel data. The model generalizes the naive credal classifier to the multilabel case. An imprecise-probabilistic quantification is achieved by means of the imprecise Dirichlet model in its global formulation. A polynomial-time algorithm to compute whether or not a label is optimal according to the maximality criterion is derived. Experimental results show the importance of robust predictions in multilabel problems.

Keywords. Credal classification, imprecise Dirichlet model, multilabel classification.

1 Introduction

A classifier represents the relationship between the characteristics of an object (*features*) and its category (*class*). A traditional *classifier* predicts the *class* variable given the value of the features. *Credal classifiers* generalize traditional classifiers, allowing for set-valued predictions of classes. A credal classifier drops the non-optimal classes returning the classes that are potentially optimal given the information available. Depending on the data, there can be one or multiple optimal classes. Credal classifiers are thus less informative but more reliable than traditional classifiers [8]. Both credal and traditional classifiers assume the classes to be mutually *exclusive.*

Multilabel classification is a modern type of classification, in which an object is allowed to have multiple *relevant* classes (or *labels*). Multilabel classification arises naturally in many domains. A news article discussing EU treaties could be labeled for instance as politics *and* finance *and* environment. Similarly, tagging of photos and videos are natural multilabel problems. In bioinformatics, the identification of the best mix of drugs for curing HIV has been addressed as a multilabel problem [14].

The simplest approach for multilabel classification is *binary relevance.* Given q labels, binary relevance develops q independent single-label classifiers. The main shortcoming of binary relevance is that it ignores the dependencies among the different classes, which in many cases are important [12]. The algorithm of classifier chain [17] is a state-of-the-art approach to model dependencies among classes. Although it achieves good empirical performance, it has no direct probabilistic interpretation.

To model the dependence among classes in a probabilistically sound way, probabilistic graphical models are typically used [1, 3, 5, 18]. Each label is represented by a Boolean variable. The i-th Boolean variable represents whether the i-th label is relevant or not for the current instance. The inference task is to detect the most probable joint configuration of the labels. A joint configuration of the labels is a *sequence* of zeros and ones. Given q labels, there are 2^q possible sequences. Evaluating the robustness of the prediction, already important in traditional classification, is even more important in multilabel classification. There is however little work on this subject.

In this paper, we tackle this problem by means of *imprecise probabilities* [19]. We propose a graphical model which generalizes the naive Bayes to the multilabel setting. We learn the model using the *imprecise Dirichlet model* (IDM) [4, 20]. We discuss two types of inferences based on the criterion of *maximality.* The joint model detects the *maximal sequences*, among the 2^q possible ones. This inference is exact but is feasible only when q is limited, for instance smaller than 10. The marginal inference detects separately the maximal states of each label. We provide an approximated algorithm to solve this inference which scales to tens of labels.

The only other example of credal multilabel classifier currently available is the recent work of Destercke [13] which devises a framework similar to binary relevance but based on credal classifiers.

The paper is organized as follows. We review some basics about Bayesian networks and the IDM in Sect. 2. We indeed show how the IDM applies to Bayesian networks in Sect. 3. The (single-label) classical naive credal classifier is reviewed in Sect. 4. The new model we present for multilabel data is described in Sect. 5.1. Classification with this model is addressed in Sect. 5.2 and the technical theorems behind the inference algorithms are in Sect. 5.3. Simulations and conclusions are in Sects. 6 and 7, while the proofs of the technical results are in the Appendix.

2 Preliminaries

We denote random variables by uppercase letters, generic values by lowercase letters and the sets of possible values by calligraphic letters. For instance X is a variable whose generic value is $x \in \mathcal{X}$. For a Boolean variable X, $\mathcal{X} := \{0, 1\}$; given a generic value $x \in \mathcal{X}$, its negation is $\neg x$.

We denote by $P(X)$ the probability mass function over X. Given a set of variables $\boldsymbol{X}$, arranged into a directed acyclic graph, a *Bayesian network* is a set of conditional tables $P(X_i|\mathrm{Pa}(X_i))$ where $\mathrm{Pa}(X_i)$ are the parents of X_i, i.e., the immediate predecessors of X_i within the graph. This defines a joint mass function $P(\boldsymbol{x}) = \prod_i P(x_i|\mathrm{pa}(X_i))$ [15].

A credal set over X is a (convex) set of probability mass functions over X. Given a credal set, the *maximality* criterion allows to choice the optimal (i.e., most probable) states as follows: $x'' \in \mathcal{X}$ is *maximal* if and only if there is no $x' \in \mathcal{X}$ s.t. $P(x') > P(x'')$ for each $P(X)$ in the credal set [19].

The *imprecise Dirichlet model* [20] (IDM) is a standard approach to learn credal sets from multinomial data. Given a variable X, a Dirichlet prior $P(\theta_x) \propto \theta^{st(x)-1}$ would induce a probability $\theta_x = \frac{n(x)+st(x)}{N+s}$. Thus, considering all the priors s.t. $\sum_x t(x) = 1$, would make θ_x to vary between $\frac{n(x)}{N+s}$ and $\frac{n(x)+s}{N+s}$.

3 IDM-based Learning with Independence

In this section we discuss the particular problem of learning a set of multivariate distributions through the IDM under specific independence assumption. This is done in the special case where the independence relations can be described within the framework of Bayesian networks. We extend Zaffalon's ideas stated in [23].

To begin the discussion let us consider the following example.

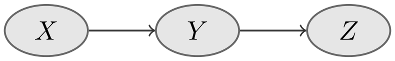


Figure 1: A chain topology

Example 1. *Consider a Bayesian network over three Boolean variables X, Y, and Z with the topology in Fig. 1. This models the conditional independence between X and Z given Y, with the joint distribution factorizing as $P(x, y, z) = P(x) \cdot P(y|x) \cdot P(z|y)$. The likelihood of a set of observations $\mathcal{D}$ is:*

$$L(\boldsymbol{\theta}) := P(\mathcal{D}|\boldsymbol{\theta}) = \prod_x \theta_x^{n(x)} \left[\prod_y \theta_{y|x}^{n(x,y)} \left[\prod_z \theta_{z|y}^{n(y,z)} \right] \right], \tag{1}$$

where $\theta_x := P(x)$, $\theta_{y|x} := P(y|x)$, and $\theta_{z|y} := P(z|y)$, for each x, y, z, and $n(\cdot)$ is the counting function. A conjugate prior over the parameters $\boldsymbol{\theta}$ is:

$$P(\boldsymbol{\theta}) \propto \prod_x \theta_x^{st(x)-1} \left[\prod_y \theta_{y|x}^{st(x,y)-1} \left[\prod_z \theta_{z|y}^{st(y,z)-1} \right] \right], \tag{2}$$

where s and the $t(\cdot)$ are nonnegative parameters. The first term in Eq. (2) is proportional to a Dirichlet prior. We set $\sum_x t(x) = 1$. Considering the corresponding (structural) constraint for the counts in the likelihood, i.e., $\sum_x n(x) = N$, we can regard s as the equivalent sample size *(ESS) of this prior distribution.*

Let us identify the additional constraints required to regard s as an ESS even for the prior in Eq. (2). We just identify the (again, structural) constraints on the likelihood $\sum_{xy} n(x, y) = \sum_{yz} n(y, z) = N$, which correspond to:

$$\sum_{xy} t(x, y) = \sum_{yz} t(y, z) = 1. \tag{3}$$

The updated parameters become therefore:

$$\theta_x = \frac{n(x) + st(x)}{N + s}, \tag{4}$$

$$\theta_{y|x} = \frac{n(x, y) + st(x, y)}{n(x) + st(x)}, \tag{5}$$

$$\theta_{z|y} = \frac{n(y, z) + st(y, z)}{n(y) + st(y)}, \tag{6}$$

with $t(x) = \sum_y t(x, y)$ and $t(y) := \sum_z t(y, z)$.

An IDM-based model is therefore obtained by considering all *the specifications of the parameters in Eqs. (4-6) consistent with the above constraints over $t(x)$,*

$t(x, y)$, and $t(y, z)$:

$$\sum_x t(x) = 1 \tag{7}$$

$$\sum_y t(x, y) = t(x), \forall x \tag{8}$$

$$\sum_z t(y, z) = \sum_x t(x, y), \forall y. \tag{9}$$

Such a model can be regarded as induced by a set of priors made of Dirichlet components and with ESS s. This is the way we generalize the IDM to multivariate models with independence. To check that the constraints are sufficient, consider all the (structural and not all independent) constraints satisfied by the count function $n(\cdot)$ in Eq. (1), i.e., $\sum_x n(x) = \sum_{xy} n(x, y) = \sum_{yz} n(y, z) = N$, $\sum_y n(x, y) = n(x)$, $\sum_z n(y, z) = n(y)$, $\sum_x n(x, y) = n(y)$. It is a trivial exercise to check that the $t(\cdot)$ parameters satisfy the analogous relations (with one replacing N).

The example deals with a node which is a child of a child of another variable. This situation does not appear in Zaffalon's original work for the naive topology, neither in other papers about more connected topologies [24].

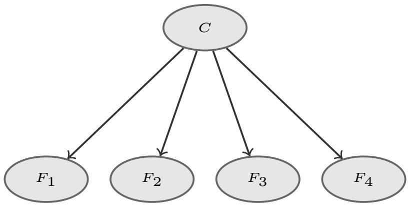
This approach can be easily extended to general Bayesian networks. The specifications over X apply to parentless nodes with Y replaced by the whole children set, the specifications over Z apply to any childless node with Y replaced by the whole parents set, and those for Y apply to any non-root non-leaf node with the parents and children playing the role of X and Z.

This section provides guidelines for learning the parameters of Bayesian networks based on the IDM. The resulting model is a *credal network* [9], with the local parameters taking their values from different credal sets, but with the constraints over the parameters of the prior inducing a *non-separate* specification [2].

4 The Naive Credal Classifier

In this section we briefly review the credal version of the naive Bayes classifier as proposed by Zaffalon in [23]. We denote the class variable as C and the feature variables as $\boldsymbol{F} := (F_1, \ldots, F_m)$. A dataset of N complete i.i.d. joint observations of $(C, \boldsymbol{F})$ is available together with a counting function $n(\cdot)$.

The features are assumed to be conditionally independent given the class. This corresponds to the topology in Fig. 6 and induces the factorization $P(c, \boldsymbol{f}) = P(c) \cdot \prod_{i=1}^m P(f_i|c)$, for each $c \in \mathcal{C}$ and $\boldsymbol{f} := (f_1, \ldots, f_m) \in \prod_{i=1}^m \mathcal{F}_i$.



By proceeding as in Ex. 1, we have:

$$P(c) = \frac{n(c) + st(c)}{N + s}, \tag{10}$$

$$P(f_i|c) = \frac{n(c, f_i) + st(c, f_i)}{n(c) + st(c)}, \tag{11}$$

for each $f_i \in \mathcal{F}_i$, $c \in \mathcal{C}$, $i = 1, \ldots, m$. The class labels assigned to an unannotated instance $\boldsymbol{f}$ of the features are those s.t. $\arg\max_{c \in \mathcal{C}} P(c, \boldsymbol{f})$.

The IDM constraints on the above positive parameters are: $\sum_c t(c) = 1$ and $\sum_{f_i} t(c, f_i) = t(c)$, for each $i = 1, \ldots, m$ and $c \in \mathcal{C}$.[1] We denote as $\boldsymbol{t}$ a generic value for the joint variable of these parameters and by $\mathcal{T}$ the corresponding feasible region.

The class labels assigned to $\boldsymbol{f}$ by this credal classifier are the *undominated* ones according to the maximality criterion. Given $c', c'' \in \mathcal{C}$, c' dominates c'' if $P(c', \boldsymbol{f}) > P(c'', \boldsymbol{f})$ for any specification consistent with the IDM constraints. This is equivalent to check:

$$\inf_{\boldsymbol{t} \in \mathcal{T}} \left[\frac{n(c'') + st(c'')}{n(c') + st(c')} \right]^{m-1} \prod_{i=1}^m \frac{n(c', f_i) + st(c', f_i)}{n(c'', f_i) + st(c'', f_i)} > 1. \tag{12}$$

The optimization of the second term can be achieved independently. The objective function rewrites as:

$$\left[\frac{n(c'') + st(c'')}{n(c') + st(c')} \right]^{m-1} \prod_{i=1}^m \frac{n(c', f_i)}{n(c'', f_i) + st(c'')}, \tag{13}$$

with the constraints being simply now $t(c') + t(c'') = 1$, with $t(c'), t(c'') > 0$. In other words, we can express the objective function as a function of a single variable. Its logarithmic derivative is a linear fractional variable, and the second derivative is always positive. Overall the minimization can be efficiently achieved by bracketing (see [23] for the details).

5 The Multilabel Credal Classifier

5.1 Model Specification

In this section we extend the setup of the previous section to multilabel classification. The class

[1] The strict positivity is required because otherwise the corresponding prior would be improper.

variable C is replaced by q (Boolean) class labels $\boldsymbol{C} := (C_1, \ldots, C_q)$, where q is the cardinality of $\mathcal{C}$. This is standard way to cope with non-exclusivity: if the j-th label of $\mathcal{C}$ is active $C_j = 1$, otherwise $C_j = 0$.

We call C_1 the *superclass*, and the other class labels *subclasses*. We assume the conditional independence of the subclasses given the superclass. Simplistically we set as superclass the class which is more frequently observed as active. The dependencies between classes can be learned in more sophisticated way, optimizing for instance the Bayesian scores [7] of the graph which connects the classes.

A dataset of N joint observations of $(\boldsymbol{C}, \boldsymbol{F})$ is available together with a counting function $n(\cdot)$.

Each feature is *replicated* q times. For each $k = 1, \ldots, m$, $\{F_k^j\}_{j=1}^q$ are replicas of F_k. For each $j = 1, \ldots, q$, the replicated features $\{F_k^j\}_{k=1}^m$ are assumed to be independent given C_j. This is a simplifying assumption, already formulated in other papers [3]. Strictly speaking, an additional dummy child modeling the fact that all the replicas corresponds to the same variable should have been added.

Accordingly, the joint factorizes as follows:

$$P(\boldsymbol{c}, \boldsymbol{f}) = P(c_1) \left[\prod_{i=2}^{q} P(c_i|c_1) \right] \prod_{j=1}^{q} \prod_{k=1}^{m} P(f_k^j|c_j), \tag{14}$$

where the values of the class labels and of the features are those consistent with $\boldsymbol{c}$ and $\boldsymbol{f}$. Parameters in Eq. (14) can be learned from the data through a procedure similar to that in the previous sections, i.e.,

$$P(c_1) = \frac{n(c_1) + st(c_1)}{n + s}, \tag{15}$$

$$P(c_i|c_1) = \frac{n(c_1, c_i) + st(c_1, c_i)}{n(c_1) + st(c_1)}, \tag{16}$$

$$P(f_k^j|c_j) = \frac{n(c_j, f_k) + st(c_j, f_k^j)}{n(c_j) + st(c_j)}. \tag{17}$$

An IDM-like version is obtained by considering all the models consistent with the following constraints:[2]

$$\sum_{c_1} t(c_1) = 1, \tag{18}$$

$$\sum_{c_i} t(c_1, c_i) = t(c_1), \forall c_i \tag{19}$$

$$\sum_{f_k^j} t(c_j, f_k^j) = \sum_{c_1} t(c_1, c_j) = t(c_j), \forall c_j, \tag{20}$$

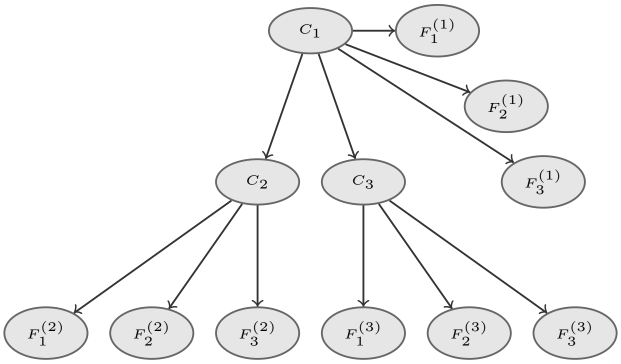


Figure 2: The multilabel naive topology

together with the strict positivity of all the parameters. Even in this case we denote by $\boldsymbol{t}$ the generic value of the joint variable including all these parameters and by $\boldsymbol{\mathcal{T}}$ the corresponding feasible region. The imprecision in this model can be regarded as induced by s missing observations, which we are completely ignorant about.

5.2 Maximal Sequences and Maximal Labels

Consider a complete observation $\boldsymbol{f}$ of the features and two sequences of labels $\boldsymbol{c}'$ and $\boldsymbol{c}''$. According to maximality, the second sequence is undominated by the first if and only if there is (at least) a prior consistent with the constraints s.t. the first sequence is less (or equally) probable than the second, i.e.,[3]

$$\inf_{\boldsymbol{t} \in \boldsymbol{\mathcal{T}}} \frac{P_{\boldsymbol{t}}(\boldsymbol{c}', \boldsymbol{f})}{P_{\boldsymbol{t}}(\boldsymbol{c}'', \boldsymbol{f})} \leq 1. \tag{21}$$

In Section 5.3 we discuss how to ascertain whether sequence $\boldsymbol{c}'$ dominates $\boldsymbol{c}''$, in linear time with respect to the number of classes and features.

A more complex problem is to ascertain whether sequence $\boldsymbol{c}''$ is optimal. This happens if the condition (21) is satisfied for each possible specifications of $\boldsymbol{c}'$, i.e.,

$$\max_{\boldsymbol{c}'} \inf_{\boldsymbol{t}} \frac{P_{\boldsymbol{t}}(\boldsymbol{c}', \boldsymbol{f})}{P_{\boldsymbol{t}}(\boldsymbol{c}'', \boldsymbol{f})} \leq 1. \tag{22}$$

To detect the non-dominated sequences it is in principle necessary to compare each possible sequence $\boldsymbol{c}'$ against each possible alternative sequence $\boldsymbol{c}''$. This implies running $2^q \cdot 2^q = 2^{2q}$ tests of the same type as Eq. (21). In Section 5.3 we present a more efficient procedure, which detects the maximal sequences by running the test of Eq. (22) only once for each candidate sequence $\boldsymbol{c}''$ (i.e., 2^q times), with a substantial computational saving. We call this model the *joint* model, as it makes inference on the joint probability

[2] Here and in the following, if there is no risk of ambiguity, the arguments of the sums and the products are omitted for sake of notation. E.g., $\sum_{c_1}$ is a shortcut for $\sum_{c_1 \in \mathcal{C}_1}$.

[3] This is an alternative formulation w.r.t. that in Eq. (12).

of the labels. Yet the complexity of the joint is exponential in the number of labels; thus the identification of the optimal sequences is feasible only of the number of classes is limited, for instance $q < 10$.

We thus devise a different approach in order to deal with datasets containing many labels. It looks for the maximal states of *each label* rather than for the maximal sequences. We call this approach the *marginal* model. The marginal inference has polynomial complexity (see Section 5.3); it is however less informative than the detection of the maximal sequences. Consider having detected k labels whose maximal states are both *relevant* and *non-relevant.* The 2^k sequences obtained combining their states in all possible ways contain the maximal sequences and others non-maximal sequences. It is not possible to know which of the 2^k sequences is maximal and which is non-maximal.

This approach corresponds to the following optimization task:

$$\min_{\boldsymbol{c}'':c''_l=1} \max_{\boldsymbol{c}'} \inf_{\boldsymbol{t}} \frac{P_{\boldsymbol{t}}(\boldsymbol{c}',\boldsymbol{f})}{P_{\boldsymbol{t}}(\boldsymbol{c}'',\boldsymbol{f})} \leq 1, \tag{23}$$

for each $l = 1, \ldots, q$, with the minimum over all the specifications of the second sequence s.t. $c''_l = 1$. If the inequality is satisfied, then there is at least an optimal sequence whose l-th label is active. By replacing $c''_l = 1$ with $c''_l = 0$, we can decide if there is an optimal sequence with the l-th label inactive.[4]

By iterating the test in Eq. (23) and its analogous with $c_l = 0$ for each $l = 1, \ldots, q$, we can decide, for each label, which one of the following three options applies: (i) all the maximal sequences have that label active; or (ii) all the maximal sequences have the label inactive; or (iii) there are maximal sequences with the label active and others with the label inactive.

We call this approach based on the joint model in Eq. (14) and the IDM constraints in Eqs. (18-20) *multi-label naive credal classifier* (MNCC). The derivation uses ideas analogous to those proposed by De Bock and de Cooman to detect the maximal sequences in hidden Markov models [11].

5.3 Solving the Optimization

In this section we present the technical results behind our implementation of the MNCC and a possible direction for its development. Let us start from the maximality-based dominance test among two sequences, which can be performed as follows.

Theorem 1. *Given two sequences $\boldsymbol{c}'$ and $\boldsymbol{c}''$ and an instance of the features $\boldsymbol{f}$, the decision task in Eq. (21) is equivalent to:*

$$\prod_{i:c'_i=\neg c''_i} \frac{n(c'_1,c'_i)\cdot g_i(c'_i,c''_i,\boldsymbol{f})}{n(c''_1,c''_i)+s} \leq 1, \tag{24}$$

if $c'_1 = c''_1$, and to

$$\inf_{0<t_1<1} h(c'_1,c''_1,t_1,\boldsymbol{f}) \prod_i \frac{n(c'_1,c'_i)\tilde{g}_i(c'_i,c''_i,\boldsymbol{f})}{n(c''_1,c''_i)+st_1}, \tag{25}$$

if $c'_1 = \neg c''_1$, where

$$g_i(c'_i,c''_i,\boldsymbol{f}) := \inf_{0<t_i<1} \prod_k \frac{\frac{n(c'_i,f_k)}{n(c'_i)+s(1-t_i)}}{\frac{n(c''_i,f_k)+st_i}{n(c''_i)+st_i}}, \tag{26}$$

$\tilde{g}_i(c'_i,c''_i,\boldsymbol{f}) := g_i(c'_i,c''_i,\boldsymbol{f})$ if $c'_i = \neg c''_i$ and one otherwise, and $h(c'_1,c''_1,t_1,\boldsymbol{f})$ is defined as

$$\left[\frac{n(c''_1)+st_1}{n(c'_1)+s(1-t_1)}\right]^{q+m-2} \prod_k \frac{n(c'_1,f_k)}{n(c''_1,f_k)+st_1}. \tag{27}$$

Furthermore, the objective functions in Eq. (25) and Eq. (26) are convex.

The proof of this theorem is in the Appendix.

Th. 1 can be used to decide whether or not $\boldsymbol{c}'$ does not dominate $\boldsymbol{c}''$. Because of the convexity results, the optima in Eq. (25) and Eq. (26) can be evaluated by bracketing (e.g., bisection) in constant time (assuming that we work with finite precision). Thus, the dominance test only takes $O(qf)$ time.

To detect the set of maximal sequences, the test should be iterated over all the possible pairs. Alternatively, we can adopt the approach in Eq. (22), i.e., maximizing w.r.t. $\boldsymbol{c}'$. If we add the constraint $c'_1 = c''_1$, the maximization becomes trivial because of the factorization in Eq. (24). If $\tilde{\boldsymbol{c}}'$ is the value leading to the maximum, we have $\tilde{c}'_1 = c''_1$ and, for $i > 1$,

$$\tilde{c}'_i := \begin{cases} \neg c''_i & \text{if} \quad \frac{n(c''_1,\neg c''_i)g_i(\neg c''_i,c''_i,\boldsymbol{f})}{n(c''_1,c''_i)+s} > 1, \\ c''_i & \text{otherwise.} \end{cases} \tag{28}$$

Thus, we perform the dominance test as in Th. 1 with $\tilde{\boldsymbol{c}}'$ and $\boldsymbol{c}''$. We similarly proceed for $c'_1 = \neg c''_1$ by considering Eq. (25) instead of Eq. (24). If t^*_1 is the value leading to the infimum, the task rewrites as:

$$\max_{c'_2,\ldots,c'_q} \left[h(\neg c''_1,c''_1,t^*_1,\boldsymbol{f}) \prod_i \frac{n(\neg c''_1,c'_i)\tilde{g}_i(c'_i,c''_i,\boldsymbol{f})}{n(c''_1,c''_i)+st^*_1} \right]. \tag{29}$$

[4] By removing the constraints $c''_l = 1$ from Eq. (23) we test whether there is a maximal sequence. But this is true by definition. Thus, if the inequality in Eq. (23) is not satisfied for $c''_l = 1$, then it should be satisfied for $c''_l = 0$, and vice versa.

The value of t_1^* depends on $\boldsymbol{c}'$ and the maximization cannot be distributed over the product as in the previous case. Nevertheless, for the i-th term of the product, a maximization w.r.t. $c_i' \in \{\neg c_i'', c_i''\}$ would be:

$$\max \left\{ \frac{n(\neg c_1'', \neg c_i'') g_i(\neg c_i'', c_i'', \boldsymbol{f})}{n(c_1'', c_i'') + st_1^*}, \frac{n(\neg c_1'', c_i'')}{n(c_1'', c_i'') + st_1^{**}} \right\}, \tag{30}$$

with the double star denoting the fact that the two optima w.r.t. t_1 can be different. Sufficient conditions for one of these two terms being the maximum irrespectively of the values of t_1^* and t_1^{**} can be used to determine $\tilde{\boldsymbol{c}}'$ as in the previous case, i.e.,

$$\tilde{c}_i' := \begin{cases} \neg c_i'' & \text{if } \frac{n(\neg c_1'', \neg c_i'') g_i(\neg c_i'', c_i'', \boldsymbol{f})}{n(c_1'', c_i'') + s} > \frac{n(\neg c_1'', c_i'')}{n(c_1'', c_i'')}, \\ c_i'' & \text{if } \frac{n(\neg c_1'', \neg c_i'') g_i(\neg c_i'', c_i'', \boldsymbol{f})}{n(c_1'', c_i'')} < \frac{n(\neg c_1'', c_i'')}{n(c_1'', c_i'') + s}. \end{cases} \tag{31}$$

Yet, unlike the specification in Eq. (28), it might be that none of the two inequalities in Eq. (31) are satisfied, and the corresponding value of $\tilde{c}_i'$ remains undefined. If this is the case, we heuristically set the value of $\tilde{c}_i'$ corresponding to the limit of Eq. (31) for small values of $s > 0$.[5]

The above approach, whose complexity is the same as a single dominance test, i.e., $O(qf)$, can be used to decide whether or not a sequence $\boldsymbol{c}''$ is maximal. This is the case if the test in Th. 1 is satisfied for both the specifications of $\boldsymbol{c}'$ in Eq. (28) and Eq. (31).

To obtain the whole set of optimal sequences, we iterate this procedure over all the 2^q possible specifications of $\boldsymbol{c}''$. To avoid this exponential blow-up, the approach in Eq. (23), i.e., minimizing w.r.t. $\boldsymbol{c}''$ with a fixed value for c_l'', can be considered instead. In practice this corresponds to minimize the maximum between the above considered expressions for $c_1' = c_1''$ and $c_1' = \neg c_1''$. Although each one of the two expressions factorizes, moving the minimum w.r.t. the different factors inside the two arguments of the maximum might introduce an approximation, i.e.,

$$\min_{c_1''} \min_{c_2'', \ldots, c_q''} \max_{c_1'} \max_{c_2', \ldots, c_2'} \inf_{\boldsymbol{t}} \frac{P_{\boldsymbol{t}}(\boldsymbol{c}', \boldsymbol{f})}{P_{\boldsymbol{t}}(\boldsymbol{c}'', \boldsymbol{f})} \geq$$

$$\min_{c_1''} \max_{c_1'} \min_{c_2'', \ldots, c_q''} \max_{c_2', \ldots, c_2'} \inf_{\boldsymbol{t}} \frac{P_{\boldsymbol{t}}(\boldsymbol{c}', \boldsymbol{f})}{P_{\boldsymbol{t}}(\boldsymbol{c}'', \boldsymbol{f})}, \tag{32}$$

where the constraint $c_l'' = 1$ on both sides is left implicit for sake of readability. The above inequality trivially follows from the technical result here below.

Lemma 1. *Given two arrays $\vec{a}$ and $\vec{b}$ with the same length n, the following inequality holds:*

$$\min_i \max\{a_i, b_i\} \geq \max\{\min_i a_i, \min_i b_i\} \tag{33}$$

where a_i and b_i are the i-th elements of $\vec{a}$ and $\vec{b}$, and the minima are intended w.r.t. $i = 1, \ldots, n$.

The proof of this lemma is in the Appendix. The right-hand side of Eq. (32) can be efficiently evaluated by reducing it to a single dominance test as we did in the first part of this section for the task in Eq. (22). If its value is (strictly) greater than one, Eq. (32) implies that also the left-hand side of Eq. (23) is greater than one, i.e., there is no maximal sequence with the l-th label active. If this is the case, we conclude that *all* the maximal sequences have the l-th label inactive. If the analogous optimization with the constraint $c_l'' = 0$ instead of $c_l'' = 1$ gives a result greater than one, we similarly conclude that all the maximal sequences have the l-th label active. Finally, if none of the above two is the case, we adopt a cautious approach by stating that there could either be maximal sequences with the l-th label active and inactive. The above approach can be considered to efficiently characterize the set of maximal sequences of the MNCC by means of an outer approximation.

6 Experiments

We compare the two variants of MNCC (joint model and marginal model) with the Bayesian graphical model, whose structure is as in Fig. . We adopt the BDeu prior [15, Chap.17] to learn the Bayesian model. This model is referred to in the following as the Bayesian model.

We consider four benchmark datasets, whose characteristics are reported in Tab. 1. *Emotions*, *Scene*, and *Slashdot* are classical benchmark datasets for multilabel classifiers. The *E-mobility* dataset is taken from a mobility study. It tracks which means of transport (car, train, bus, etc.) are used by a person for a given trip. The features are constituted by the length and duration of the trip, hour and day of the week, number of persons, reason of the trip, etc. [6].

Data set	Classes	Features	Instances
Emotions	6	44/72	593
Scene	6	224/294	2407
E-mobility	10	14/18	4226
Slashdot	22	496/1079	3782

Table 1: Benchmark datasets.

We validate the classifiers by a ten-folds cross-validation. Before training any classifier, we perform two pre-processing steps. First, we discretize numerical features into four bins. Then we perform feature selection as follows. We adopt the correlation-based feature selection (CFS) [21, Chap. 7.1], often used in

[5] If $n(\neg c_1'', \neg c_i'') g_i(\neg c_i'', c_i'', \boldsymbol{f}) \neq n(\neg c_1'', c_i'')$, it is easy to check that the two inequalities cannot be simultaneously satisfied and, for sufficiently small s, one of them is always satisfied.

traditional classification. We perform CFS q times, once for each different label. Eventually, we retain the *union* of the features selected in the q runs. This is a useful pre-processing step which reduces the number of features, removing the non-relevant ones. As an example, Tab. 1 displays the number of features after and before this selection procedure when applied to the benchmark datasets considered in this paper. Feature selection for multilabel classification is however an open problem, and more sophisticated approaches can be designed to this end.

We start by assessing the joint model. We measure the *exact match* of the Bayesian model, namely the proportion of times in which the whole sequence of classes has been correctly predicted. For the MNCC we measure the *# of sequences*, namely the number of maximal sequences; moreover we measure the *credal match*, namely the proportion of times in which the actual sequence belongs to the set of optimal sequences.

Dataset	Bayesian	Credal (MNCC)	
	Exact match	# of seqs	Credal match
Emotions	.27	9.4	.80
Scene	.29	7.6	.80

Table 2: Experimental results of the joint model.

The sequence predicted by the Bayesian model is always recognized as maximal. The credal joint model is more robust than its Bayesian counterpart: the credal match is about three times larger than the total accuracy of the Bayesian multilabel classifier (see Tab.2). The number of maximal sequences is reasonably limited, considering that the presence of 6 classes implies 64 possible sequences. The exact match of the Bayesian classifier drops sharply on the instances which have many maximal sequences. On the Scene dataset, the total accuracy is 0.23 and 0.40 on the instances which have respectively less and more than nine maximal sequences. A similar pattern is observed also on the Emotions dataset. These results are obtained through the joint model, which enumerates all the 2^q possible sequences and checks whether they are maximal as in Eq. (22). They show the interesting potential of the credal approach to multilabel classification. Yet, the joint model can only cope with small q.

The marginal model can deal with larger q and thus can be tested on more challenging datasets. We adopt the outer approximation corresponding to the dominance test in Eq. (23). Results of a ten-folds cross validations are in Figs. 3–5. We evaluate the marginal model label-wise. In particular we measure for each

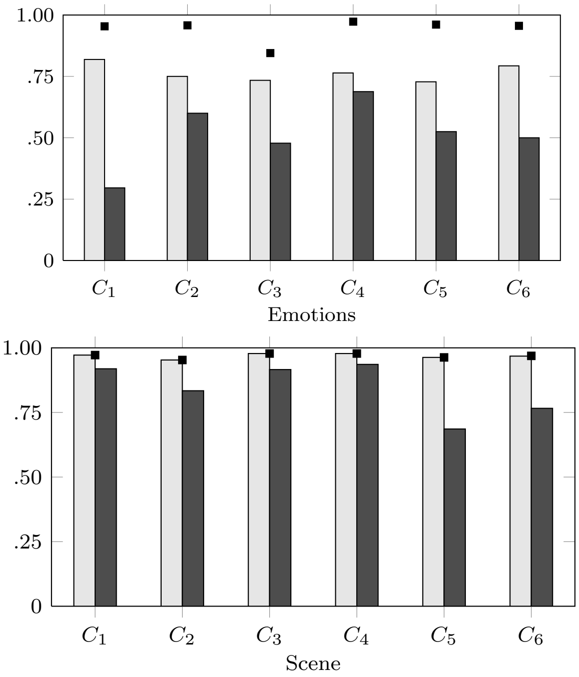


Figure 3: Accuracy of the Bayesian model on the instances on which the marginal MNCC model is determinate (light bars) and indeterminate (dark bars). The black squares denote the determinacy level. The results are presented label-wise.

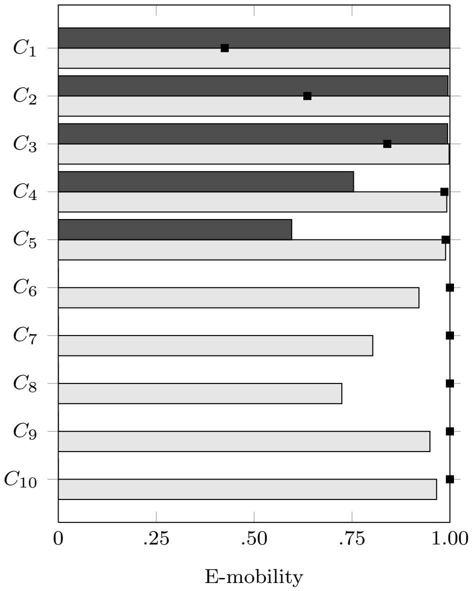


Figure 4: Accuracy of the Bayesian model on the E-mobility dataset. Light gray bars denote the accuracy when the marginal MNCC model is determinate. When determinacy (black squares) is one, the dark gray bar associated to the case when MNCC is indeterminate is not shown. The results are presented label-wise.

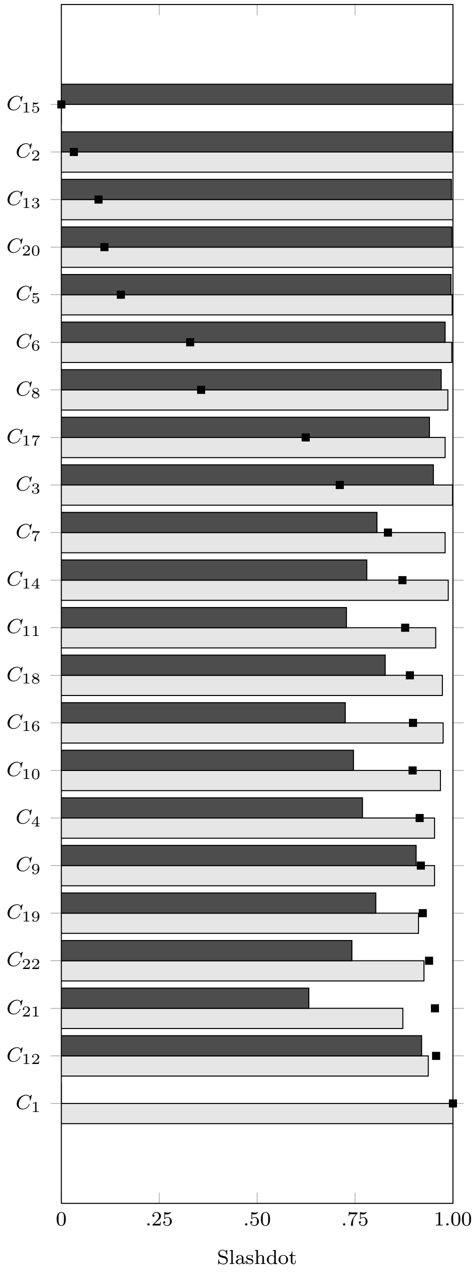


Figure 5: Accuracy of the Bayesian model on the Slashdot dataset. The dark gray bars denote the accuracy of the Bayesian model when the MNCC is indeterminate. If the determinacy (black squares) is zero, the light gray bar corresponding to the cases when the MNCC is determinate is undefined. Labels are sorted according to the determinacy level just for sake of readability. The results are presented label-wise.

label the accuracy of Bayesian model when MNCC returns a determinate and an indeterminate prediction. We also report the *determinacy*, i.e. the proportion of instances on which MNCC is determinate. On Scene and Emotions the accuracy of the Bayesian model sharply drops when the multilabel classifier becomes indeterminate. This confirms a well-known strength of credal classifiers compared to Bayesian classifiers [8]. This is generally confirmed also on E-mobility and Slashdot. However in these datasets there are also labels in which the Bayesian model is perfectly accurate when the credal model is indeterminate (see the first labels of both datasets). This suggests that the credal model is excessively indeterminate in some situations. This is a problem which is also known in traditional classification and which could be mitigated for instance by ϵ-contaminating the IDM with the uniform prior.

Future studies might inspect also further indicator of performance for multilabel classification, such as the F-metric. We focus on the exact match and on the label-wise accuracy as the inferences for this indicators are optimal. Optimal inferences for other indicators have still to be developed.

A Matlab software implementation of the MNCC is freely available at `http://ipg.idsia.ch/software`.

7 Conclusions

We have generalized the naive credal classifier to cope with multilabel data. The preliminary experiments are promising: the credal approach yields more robust predictions than the Bayesian approach. To scale to large number of labels it is necessary adopting the marginal model, whose inference is approximated.

As future work, it could be interesting to compare the inferences yielded by local and the global specification of the IDM (e.g., by exploiting some of the results in [10]). Moreover one could consider optimality criteria others than maximality (e.g., E-admissibility). A comparison with other methods possibly yielding multiple sequences (e.g., [16, 22]) could be also considered.

Acknowledgements

We thank Claudio Bonesana for support during the preparation of the datasets. We thank Jasper De Bock and Cassio de Campos for stimulating discussions about MAP tasks in credal networks. We also thank Denis Mauá for his suggestions about possible justification of models with the replicated features.

References

[1] A. Antonucci, G. Corani, D.D. Mauá, and S. Gabaglio. An ensemble of Bayesian networks for multilabel classification. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI-13)*, pages 1220–1225, 2013.

[2] A. Antonucci and M. Zaffalon. Decision-theoretic specification of credal networks: a unified language for uncertain modeling with sets of Bayesian networks. *International Journal of Approximate Reasoning*, 49(2):345–361, 2008.

[3] J. Arias, J. Gámez, T.D. Nielsen, and J.M. Puerta. A pairwise class interaction framework for multilabel classification. In L. van der Gaag and A. Feelders, editors, *PGM'14: Proceedings of the Seventh European Workshop on Probabilistic Graphical Models*, Lecture Notes in Artificial Intelligence, pages 17–32. Springer, 2014.

[4] J.M. Bernard. An introduction to the imprecise dirichlet model for multinomial data. *International Journal of Approximate Reasoning*, 39(2):123–150, 2005.

[5] C. Bielza, G. Li, and P. Larrañaga. Multi-dimensional classification with Bayesian networks. *International Journal of Approximate Reasoning*, 52(6):705–727, 2011.

[6] F. Cellina, A. Frster, D. Rivola, L. Pampuri, R. Rudel, and A. Rizzoli. Using smartphones to profile mobility patterns in a living lab for the transition to e-mobility. In J. Hebiek, G. Schimak, M. Kubasek, and A. Rizzoli, editors, *Environmental Software Systems. Fostering Information Sharing*, volume 413 of *IFIP Advances in Information and Communication Technology*, pages 154–163. Springer, 2013.

[7] G. Corani, A. Antonucci, D. Mauá, and S. Gabaglio. Trading off Speed and Accuracy in Multilabel Classification. In L. van der Gaag and A. Feelders, editors, *PGM'14: Proceedings of the Seventh European Workshop on Probabilistic Graphical Models*, Lecture Notes in Artificial Intelligence, pages 145–159. Springer, 2014.

[8] G. Corani and M. Zaffalon. Learning reliable classifiers from small or incomplete data sets: The naive credal classifier 2. *The Journal of Machine Learning Research*, 9:581–621, 2008.

[9] F. G. Cozman. Credal networks. *Artificial Intelligence*, 120:199–233, 2000.

[10] J. De Bock, C.P. de Campos, and A. Antonucci. Global sensitivity analysis for MAP inference in graphical models. In *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, 2014.

[11] J. De Bock and G. de Cooman. An efficient algorithm for estimating state sequences in imprecise hidden Markov models. *Journal of Artificial Intelligence Research*, 50:189–233, 2014.

[12] K. Dembczynski, W. Waegeman, and E. Hüllermeier. An analysis of chaining in multi-label classification. In *Proceedings of the 20th European Conference on Artificial Intelligence (ECAI)*, pages 294–299, 2012.

[13] S. Destercke. Multilabel predictions with sets of probabilities: the Hamming and ranking loss cases. *Pattern Recognition*, 2015.

[14] D. Heider, R. Senge, W. Cheng, and E. Hüllermeier. Multilabel classification for exploiting cross-resistance information in HIV-1 drug resistance prediction. *Bioinformatics*, 29(16):1946–1952, 2013.

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

[16] I. Pillai, G. Fumera, and F. Roli. Multi-label classification with a reject option. *Pattern Recognition*, 46(8):2256–2266, 2013.

[17] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359, 2011.

[18] L.C. Van Der Gaag and P.R. De Waal. Multi-dimensional Bayesian network classifiers. In M. Studený and J. Vomlel, editors, *Proc. of the 3rd European Workshop on Probabilistic Graphical Models (PGM '06)*, pages 107–114. Action M, 2006.

[19] P. Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, New York, 1991.

[20] P. Walley. Inferences from multinomial data: Learning about a bag of marbles. *Journal of the Royal Statistical Society: Series B*, 58:3–34, 1996.

[21] I.H. Witten, E. Frank, and M.A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, 2011.

[22] Z. Younes, F. Abdallah, and T. Denoeux. Fuzzy multi-label learning under veristic variables. In *Proceedings of the IEEE International Conference on Fuzzy Systems*, pages 1–8. IEEE, 2010.

[23] M. Zaffalon. Statistical inference of the naive credal classifier. In G. de Cooman, T.L. Fine, and T. Seidenfeld, editors, *ISIPTA '01: Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications*, pages 384–393, The Netherlands, 2001. Shaker.

[24] M. Zaffalon and E. Fagiuoli. Tree-based credal networks for classification. *Reliable Computing*, 9(6):487–509, 2003.

A Proofs

Proof of Theorem 1. *We consider the objective function in Eq.* (21) *by distinguishing whether or not the two sequences* $\boldsymbol{c}'$ *and* $\boldsymbol{c}''$ *share the first label, i.e.,*

$$\frac{P_{\boldsymbol{t}}(\boldsymbol{c}',\boldsymbol{f})}{P_{\boldsymbol{t}}(\boldsymbol{c}'',\boldsymbol{f})}=\begin{cases} G_{\boldsymbol{t}}(\boldsymbol{c}',\boldsymbol{c}'',\boldsymbol{f}), & \text{if } c_1'=c_1'', \\ \\ H_{\boldsymbol{t}}(\boldsymbol{c}',\boldsymbol{c}'',\boldsymbol{f}), & \text{if } c_1'=\neg c_1''. \end{cases} \tag{34}$$

Because of Eq. (14), *function* $G_{\boldsymbol{t}}(\boldsymbol{c}',\boldsymbol{c}'',\boldsymbol{f})$ *writes as:*

$$\prod_{i:c_i'=\neg c_i''}\left[\frac{n(c_1',c_i')+st(c_1',c_i')}{n(c_1'',c_i'')+st(c_1'',c_i'')}\prod_{k=1}^{m}\frac{\frac{n(c_i',f_k)+st(c_i',f_k)}{n(c_i')+st(c_i')}}{\frac{n(c_i'',f_k)+st(c_i'',f_k)}{n(c_i'')+st(c_i'')}}\right], \tag{35}$$

where the restriction in the outer product is possible because of the contribution of the other terms is one (remember that $c_1'=c_1''$*). A preliminary optimization w.r.t. the constraints can be achieved as in Sect.* 4 *by setting* $t(c_i',f_k)\to 0$ *and* $t(c_i'',f_k)\to t(c_i'')$ *(remember that* $c_i'=\neg c_i''$*). Similarly,* $t(c_1',c_i')\to 0$ *and* $t(c_1'',c_i'')\to t(c_1'')$*. After these operations, the result rewrites as:*

$$\prod_i{}'\left[\frac{n(c_1',c_i')}{n(c_1'',c_i'')+st(c_1'')}\prod_k\frac{\frac{n(c_i',f_k)}{n(c_i')+st(c_i')}}{\frac{n(c_i'',f_k)+st(c_i'')}{n(c_i'')+st(c_i'')}}\right], \tag{36}$$

where the prime in the product is a shortcut for the restriction. The optimization w.r.t. $t(c_1'')$ *is achieved in the limit* $t(c_1'')\to 1$*. Even the remaining optimization tasks can be achieved independently of the others. The result is the left-hand side of Eq.* (24)*, where, in Eq.* (26)*, we have set* $t_i:=t(c_i'')$*, and hence* $t(c_i')=1-t_i$ *(remember that, for these terms,* $c_i'=\neg c_i''$*).*

We similarly proceed for $H_{\boldsymbol{t}}(\boldsymbol{c}',\boldsymbol{c}'',\boldsymbol{f})$*, i.e., because of Eq.* (34) *and Eq.* (14)*:*

$$\left[\frac{n(c_1'')+st(c_1'')}{n(c_1')+st(c_1')}\right]^{q+m-2}\prod_k\frac{n(c_1',f_k)+st(c_1',f_k)}{n(c_1'',f_k)+st(c_1'',f_k)}$$
$$\prod_i\frac{n(c_1',c_i')+st(c_1',c_i')}{n(c_1'',c_i'')+st(c_1'',c_i'')}\prod_j{}'\prod_k\frac{\frac{n(c_j',f_k)+st(c_j',f_k)}{n(c_j')+st(c_j')}}{\frac{n(c_j'',f_k)+st(c_j'',f_k)}{n(c_j'')+st(c_j'')}}. \tag{37}$$

As in the previous case, we perform some optimization, rename the remaining variables, and independently optimize w.r.t. t_i $(i>1)$*. Afterwards, we optimize w.r.t.* t_1 *and* $\inf_{\boldsymbol{t}} H_{\boldsymbol{t}}(\boldsymbol{c}',\boldsymbol{c}'',\boldsymbol{f})$ *becomes as in Eq.* (25)*.*

Finally, we prove that the objective functions in the right-hand side of Eq. (26) *and in Eq.* (25) *are convex. The derivative of the logarithm of the objective function in the right-hand side of Eq.* (26) *divided by the positive constant* s *is equal to:*

$$\frac{m}{n(c_i')+s(1-t_i)}-\sum_k\frac{1}{n(c_i'',f_k)+st_i}+\frac{m}{n(c_i)''+st_i}. \tag{38}$$

The second derivative, again divided by s*, is:*

$$\frac{m}{[n(c_i')+s(1-t_i)]^2}+\sum_k\frac{1}{[n(c_i'',f_k)+st_i]^2} \tag{39}$$
$$-\frac{m}{[n(c_i)''+st_i]^2}, \tag{40}$$

and its nonnegativity easily follows from $n(c_i'')\geq n(c_i'',f_k)$*. Similarly, the second derivative of the logarithm of the objective function in Eq.* (25) *is:*

$$-\frac{q+m-2}{[n(c_1'')+st_1]^2}+\frac{q+m-2}{[n(c_1')+s(1-t_1)]^2}$$
$$+\sum_k\frac{1}{[n(c_1'',f_k)+st_1]^2}+\sum_i\frac{1}{[n(c_1'',c_i'')+st_1]^2} \tag{41}$$

As in the previous case, the nonnegativity follows from $n(c_i'')\geq n(c_i'',f_k)$*.* □

Proof of Lemma 1. *We prove the result by contradiction. Thus, we assume that:*

$$\min_i\max\{a_i,b_i\}<\max\{\min_i a_i,\min_i b_i\}. \tag{42}$$

Let i^* *denote the* $\arg\min$ *of the left-hand side. If, without any lack of generality, we assume* $\min_i a_i\geq\min_i b_i$*, Eq.* (42) *rewrites as:*

$$\max\{a_{i^*},b_{i^*}\}<\min_i a_i. \tag{43}$$

If $a_{i^*}>b_{i^*}$*, we obtain the contradiction* $a_{i^*}<\min_i a_i$*. Otherwise, we have:*

$$a_{i^*}\leq b_{i^*}<\min_i a_i \tag{44}$$

which is also a contradiction. □