Subgroup Discovery meets Bayesian networks – an Exceptional Model Mining approach

Wouter Duivesteijn, Arno Knobbe
LIACS
Leiden University
The Netherlands
wouterd@liacs.nl

Ad Feelders, Matthijs van Leeuwen
ICS
Utrecht University
The Netherlands

***Abstract*—Whenever a dataset has multiple discrete target variables, we want our algorithms to consider not only the variables themselves, but also the interdependencies between them. We propose to use these interdependencies to quantify the quality of subgroups, by integrating Bayesian networks with the Exceptional Model Mining framework. Within this framework, candidate subgroups are generated. For each candidate, we fit a Bayesian network on the target variables. Then we compare the network's structure to the structure of the Bayesian network fitted on the whole dataset. To perform this comparison, we define an edit distance-based distance metric that is appropriate for Bayesian networks. We show interesting subgroups that we experimentally found with our method on datasets from music theory, semantic scene classification, biology and zoogeography.**

***Keywords*-Exceptional Model Mining, Subgroup Discovery, Bayesian networks**

I. Introduction

Exceptional Model Mining (EMM) [1] is a framework that extends traditional Subgroup Discovery. Given a single target variable and a dataset, Subgroup Discovery is concerned with finding subsets of the data where the distribution of the target variable is substantially different from its distribution on the whole dataset. EMM extends this notion to targets that are models of some sort; we try to find subsets of the data where a model fitted on the subset is substantially different from that model fitted on the whole dataset.

In this paper, we perform EMM on data with a number of discrete target variables. The interdependencies between these variables are modeled by a Bayesian network. The method we introduce strives to find subgroups that balance two properties. On the one hand, we seek subgroups that have a high degree to which the interdependencies between targets in the subgroup differ from those in the whole data. To quantify this degree, we define a distance metric on Bayesian networks, loosely based on edit distance [2]. On the other hand, we seek subgroups with a size that is not extreme: subgroups that consist of only a few data points or cover nearly the whole dataset are not interesting.

One possible application of our method is in the field of multi-label classification (MLC) [3]. MLC is a generalization of traditional classification: each instance is allowed to be a member of precisely one class in classification, but any number of classes in MLC. When viewing MLC as an example of our method, the presence or absence of a certain label corresponds to one target variable. Notice however that the scope of the proposed method is wider: the targets need not at all be labels attached to a data point.

To illustrate the use of interdependencies in a Bayesian network as a relatively complex target concept, we consider the research performed by Robert T. Paine in 1963 and 1964 in Makah Bay, Washington [4]. It concerns the carnivore starfish *Pisaster ochraceus* whose presence kept a marine ecosystem consisting of 15 species stable. In this system, the sponge *Haliclona* was browsed upon by the nudibranch *Anisodoris*. When *Pisaster* was artificially removed, the bivalve *Mytilus californianus* and the barnacles *Balanus glandula* and *Mitella polymerus* rapidly grew and crowded out other species. In total, only 8 species remained. Also, the sponge-nudibranch food chain was displaced, and the anemone population was reduced in density. When present, *Pisaster* does not eat either of these carnivores or the sponge.

Paine remarks that the food chains are strongly influenced by *Pisaster*, but by an indirect process. When dealing with a dataset detailing the presence of individual species, existing methods can probably detect simple patterns in the ecosystem, such as the growth of *Mytilus*, *Balanus* and *Mitella* and the decline in the number of species when *Pisaster* is removed. However, the more indirect influence of *Pisaster* on processes such as a food chain it is not direcly related to, like the one between *Haliclona* and *Anisodoris*, cannot be found by looking at single species or even correlations between pairs of species: the (in-)dependence between *Haliclona* and *Anisodoris* is conditional on the presence of *Pisaster*. Our Bayesian network approach enables the consideration of conditional dependencies, thus making detection possible of indirect processes that can be captured with a Bayesian network. For instance, in the marine biology example we can find a subgroup defined by environmental parameters in which complex food chains are displaced. The ability to cope with Bayesian networks makes our method applicable

to datasets not only from marine ecosystems, but also from such diverse fields as traffic accident reconstruction [5], medical expert systems [6], and financial operational risk [7].

The main contributions of this paper are the definition of a tractable distance metric on the structures of Bayesian networks with the same set of vertices, and an Exceptional Model Mining method for finding interesting subgroups that explicitly employs the interdependencies between discrete variables. We specifically discuss the computational complexity of our metric, since in any EMM process a large number of candidate subsets of the data is considered, and it is hence essential that the methods that we integrate with the EMM process do not impose a heavy computational burden.

This paper is organized as follows. We introduce required notation in Section II. In Section II-A the EMM framework is reiterated. Section III contains our novel method to apply the EMM framework to data with multiple discrete target variables. In Section IV we discuss experimental results with several multi-target datasets. Section V concludes the paper with a summary, and some pointers for further research.

II. Preliminaries

Throughout this paper, we assume a dataset D with elements (*data points*) $\overrightarrow{x} \in D$ of the form $\overrightarrow{x} = \{a_1, \ldots, a_k, t_1, \ldots, t_m\}$, where k and m are positive integers. The set $\{a_1, \ldots, a_k\}$ is denoted by $\overrightarrow{a}$, which we call the vector of *attributes* of $\overrightarrow{x}$, and the set $\{t_1, \ldots, t_m\}$ is denoted by $\overrightarrow{t}$, which we call the vector of *targets* of $\overrightarrow{x}$. Each target t_i is assumed to be discrete, and each vector of attributes is taken from an unspecified domain $\mathcal{A}$. We refer to the ith data point by $\overrightarrow{x^i}$, its attributes by $\overrightarrow{a^i}$, and its jth target by t^i_j. We omit the superscript if no confusion can arise. The size of the dataset is denoted by $N = |D|$.

For our definition of subgroups we need to define *patterns*. These are functions $p : \mathcal{A} \to \{0, 1\}$. A pattern p *covers* a data point $\overrightarrow{x^i}$ if and only if $p\left(\overrightarrow{a^i}\right) = 1$.

Definition (Subgroup). A *subgroup* corresponding to a pattern p is the bag of data points $G_p \subseteq D$ that p covers:

$$G_p = \left\{ \overrightarrow{x^i} \in D \;\middle|\; p(\overrightarrow{a^i}) = 1 \right\}$$

From now on we omit the p if no confusion can arise, and refer to a subgroup as G. We write n for the size of G.

In order to objectively evaluate a candidate pattern in a given dataset, we need to define a *quality measure*. For each pattern p in the pattern language $\mathcal{P}$, this function measures how interesting the model is that we induce on G_p.

Definition (Quality Measure). A *quality measure* is a function $\varphi_D : \mathcal{P} \to \mathbb{R}$ that assigns a unique numeric value to a pattern p, given a dataset D.

A. EMM revisited

Exceptional Model Mining [1] is a data mining framework that can be seen as an extension of the Subgroup Discovery (SD) framework. SD strives to find patterns that satisfy certain user-specified constraints. Usually these constraints include lower bounds on the quality of the pattern ($\varphi(p) \geq lb_1$) and size of the induced subgroup ($n \geq lb_2$). More constraints may be imposed as the question at hand requires; domain experts may for instance request an upper bound on the complexity of the pattern. Most common SD algorithms traverse (we consider the exact search strategy to be a parameter of the algorithm) the search space of candidate patterns in a general-to-specific way: they treat the space as a lattice whose structure is defined by a *refinement operator* $\rho : \mathcal{P} \to 2^{\mathcal{P}}$. This operator determines how patterns can be extended into more complex patterns by atomic additions. Most applications (including ours) assume ρ to be a *specialization operator*: $\forall\ p_s \in \rho(p_g)\ :\ p_g \succeq p_s$ (i.e. p_s is more specialized than p_g). The algorithm results in a ranked list of patterns (or the corresponding subgroups) that satisfy the user-defined constraints.

In traditional SD $m = 1$, i.e. there is only a single target variable. Hence, the typical quality measure contains a component indicating how different the distribution over the target variable in the subgroup is, compared to its distribution in the whole dataset. Since unusual distributions are easily achieved in small subsets of the dataset, the typical quality measure also contains a component indicating the size of the subgroup. Thus, whether a pattern is deemed interesting depends on both its exceptionality and the size of the corresponding subgroup.

EMM can now be seen as an extension of SD. Rather than the regular single target variable, EMM uses a more complex target concept. For each subgroup under consideration, we induce a model on the targets $t_1, \ldots, t_m$. Then quality measures are defined that indicate how exceptional the model fitted on the targets in the subgroup is, compared to the model fitted on the targets in the whole dataset. For example, [1] proposes quality measures for correlation models, a linear regression model, and classification models.

In the EMM setting, usually the *beam search* strategy [8] is chosen, which performs a level-wise search. On each level, the best w patterns according to our quality measure φ are selected, and refined to create the candidate patterns for the next level. The search is constrained by an upper bound on the complexity of the pattern and a lower bound on the support of the corresponding subgroup. This search strategy combines the advantages of a greedy method with those of the implicit parallel search: as on each level w alternatives are considered, the search process is less likely to end up in a local optimum than a pure greedy approach, but the selection of the w best patterns at each level keeps the process focused and thus more tractable.

III. EMM IN DATA WITH MULTIPLE DISCRETE TARGET VARIABLES

In this section, we introduce our approach to use data with multiple discrete target variables in an EMM setting. In such data we ideally would look beyond the targets themselves, and take the interdependencies between the targets into account. We propose to use these interdependencies in the validation of the subgroups. In order to do so, the interdependencies need to be modeled first. We do this by fitting a Bayesian network on the target variables.

A. Bayesian networks

A *Bayesian network* [9] is a directed acyclic graph (DAG) that represents a set of random variables and the interaction effects that hold between them. Each random variable is represented by a vertex in the graph, and the edges model the independence relations between the variables by d-separation: two variable x and y are conditionally independent given a set of variables Z if x and y are d-separated relative to Z. For details on d-separation, see [9].

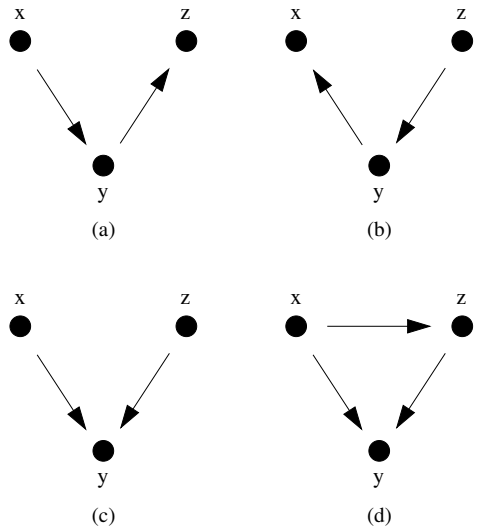


Figure 1. Example Bayesian networks

For instance, in network (a) in Figure 1, there is no path from z to x, so z and x are conditionally independent given y. This is the only independence relation in this network.

There are two important peculiarities about the independence relations in Bayesian networks. First, different Bayesian networks may represent the same independence relations. If we look at network (b), we find that in this network only one independence relation holds: x and z are conditionally independent given y. By symmetry of conditional independence, this is the same independence relation as the one in network (a). Bayesian networks that represent the same independence relations are called *equivalent*. Note that this relation partitions Bayesian networks into equivalence classes. Second, Bayesian networks with the same skeleton (the network when we drop the directions) are not necessarily equivalent. In network (c), for example, x and z are marginally independent, unlike in networks (a) and (b).

We identify a special configuration of vertices and edges in a Bayesian network that is relevant for the discussion in the rest of this paper. It is a structure as seen in network (c): a *v-structure*.

Definition (V-structure). A *v-structure* in a Bayesian network is a set of three vertices $\{x, y, z\}$ such that the network contains edges $x \rightarrow y$ and $z \rightarrow y$, but no edge between x and z.

The probabilistic interpretation of this v-structure is that x and z are marginally independent, but conditionally dependent given y. V-structures are also known as *immoralities*, since the parents of vertex y are unmarried, i.e. there is no edge between them. A graph can be *moralized* [10] by first marrying all unmarried parents (i.e. draw an edge between all pairs of vertices that have a common child but no common edge), and then dropping directions. Thus, moralizing a graph removes all v-structures. The moralized versions of the networks of Figure 1 are depicted in Figure 2. As you can see, the moralized version of network (c) has an extra edge, which corresponds to removing the v-structure in the original network.

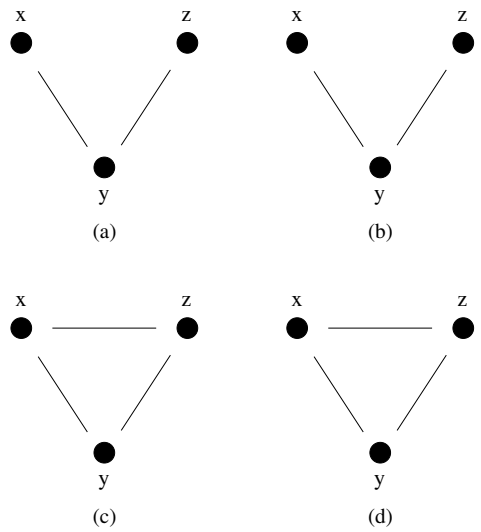


Figure 2. Moralized graphs for the networks in Figure 1

Notice that the moral graph also is not sufficient to capture all information about the underlying independency relations; x and z are marginally independent in network c and marginally dependent in network d, but these networks have the same moral graph.

B. Fitting a Bayesian network on data

There are many algorithms to learn a DAG model from data, see for instance [11]–[13]. We use a non-deterministic hill climbing algorithm; using a hill climbing method makes the algorithm speedy enough for use in an EMM setting, while its non-deterministic nature decreases the chance that the algorithm will end up in a local optimum.

We start with a Bayesian network with m vertices and no edges, and compute the quality of that model. We choose the *Bayesian Dirichlet equivalent uniform* (BDeu) score (see [14]), because it assigns equal scores to equivalent models and assumes no prior information. Then we hill climb through the space of Bayesian networks by applying the best single-edge change in the model. At each step, we apply a random number of covered arc reversals [15], in order to escape from a maximum that may be local. For more details on this combination of methods, see [16].

Notice that this process is quite non-deterministic: at every step in the hill climbing, and whenever we try to escape a maximum, a random number of randomly selected covered edges is reversed. During our experiments we occasionally found different Bayesian networks for the same data with different random seeds. However, these changes were not very dramatic: few edges change, and almost all resulting networks for the same data are equivalent.

C. Quality Measure

We use the algorithm from the previous subsection to fit a Bayesian network on the target variables restricted to the whole dataset, and restricted to candidate subgroups. Now we would like to compare the structure of these networks, in order to find the subgroups in which the interdependencies between the targets differ the most from those in the whole dataset; the exceptional models. An obvious candidate for such a comparison is edit distance [2].

Definition (Traditional edit distance). The *edit distance* between two given graphs $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ is the minimal number of edges we need to add to, remove from and reverse in E_1 to obtain E_2.

Edit distance is a common way to measure the distance between two DAGs. However, it is obvious that this does not directly translate to a sensible distance measure for Bayesian networks: as we have seen in Subsection III-A, networks a and b are equivalent, but the traditional edit distance between them is 2 instead of the desired 0. Hence, to make a suitable quality measure out of edit distance, we need to take into account the equivalence classes.

We could theoretically solve this problem via the *essential graph* [17], a partially directed, partially undirected graph that is the same for all equivalent DAGs. The essential graph of a DAG is the graph wherein we keep all directed edges that are the same in the whole equivalence class, and drop the directions of the other edges. Essential graphs can be calculated in polynomial time [17]. Unfortunately, the algorithm Andersson et al. describe takes $\mathcal{O}\left(m^6\right)$ time. By implementing the innovation from Chickering [18] this can be improved to $\mathcal{O}\left(m^5\right)$, but this is still too expensive to be used in EMM, and to our knowledge no faster algorithm exists. Instead, we propose a heuristic based on the following well-known result by Verma and Pearl [19]:

Theorem 1 (Equivalent DAGs)**.** *Two DAGs are equivalent if and only if they have the same skeleton and the same v-structures.*

Since these two conditions determine whether two DAGs are equivalent, it makes sense to consider the number of differences in skeletons and v-structures as a measure of how different two DAGs are.

Definition (Edit distance for Bayesian networks). Let two Bayesian networks BN_1 and BN_2 be given with the same set of vertices whose size we denote by m. Denote the edge set of their skeletons by S_1 and S_2, and the edge set of their moralized graphs by M_1 and M_2. Let

$$\ell = \Big|[S_1 \oplus S_2] \ \cup \ [M_1 \oplus M_2]\Big|$$

The distance between BN_1 and BN_2 is defined as:

$$d(BN_1, BN_2) = \frac{2\ell}{m(m-1)}$$

As usual in set theory, $\oplus$ denotes an exclusive disjunction: $X \oplus Y = (X \cup Y) - (X \cap Y)$. The factor $\frac{2}{m(m-1)}$ causes the distance to range between 0 and 1: it is the expanded reciprocal of $\binom{m}{2}$, the number of distinct pairs of vertices.

Notice that this distance only considers the network structures, not the underlying probability distributions; Leman et al. [1] have previously discussed a variant of EMM that assigns exceptionality to a subgroup by looking at those distributions. As a consequence, our distance does not distinguish between e.g. different signs of correlation: if an edge corresponds to a positive correlation in one network and to a negative correlation in another network, then this edge does not contribute to the distance.

We illustrate the edit distance by computing the mutual distances between the networks in Figure 1. We find that $\mathrm{d}(a, b) = 0$ and $\mathrm{d}(a, c) = \mathrm{d}(a, d) = \mathrm{d}(b, c) = \mathrm{d}(b, d) = \mathrm{d}(c, d) = 1/3$. Only for the two networks that are equivalent, distance 0 is obtained. If we compare the networks to the independence model i which has no edges at all, we obtain $\mathrm{d}(a, i) = \mathrm{d}(b, i) = 2/3$, and $\mathrm{d}(c, i) = \mathrm{d}(d, i) = 1$.

This distance can now be used to quantify the exceptionality of a subgroup:

Definition (Edit distance based quality measure). Let a pattern p be given. Denote the Bayesian network we fit on D by BN_D, and denote the Bayesian network we fit on G_p by BN_p. Then the quality of p is:

$$\varphi_{\text{ed}}(p) = d\left(BN_D, BN_p\right)$$

If we would plug φ_{ed} into the EMM framework, a problem similar to the problem with quality measures in traditional Subgroup Discovery would occur: unusual interdependencies between the targets are easily achieved in very small subsets of the dataset. Thus, using φ_{ed} would result in small subgroups. For this reason, we introduce an alternative measure that includes the size of the subgroup in the evaluation. We use the *entropy* of the split between the subgroup and the rest of the dataset to capture this [1].

Definition (Entropy).

$$\varphi_{\text{ent}}(p) = -\frac{n}{N}\lg\left(\frac{n}{N}\right) - \frac{N-n}{N}\lg\left(\frac{N-n}{N}\right)$$

Here, $\lg$ is the binary logarithm. The entropy captures the information content of the split. It favours balanced splits over skewed splits, and is again normalized to return 0 and 1 for the extreme cases (subgroup being empty or covering the whole dataset, and $50/50$ splits, respectively).

Because we do not want to find subgroups that have a low quality value on either the edit distance or the entropy measure, we make an aggregated measure.

Definition (Weighed Entropy and Edit Distance).

$$\varphi_{\text{weed}}(p) = \sqrt{\varphi_{\text{ent}}(p)} \cdot \varphi_{\text{ed}}(p)$$

The original components ranged from 0 to 1, hence the new quality measure does so too. We take the square root of the entropy, thus reducing its bias towards $50/50$ splits, since we are primarily interested in a subgroup with large edit distance, while mediocre entropy is acceptable.

In Section IV, we will focus on results obtained with φ_{weed}. In addition, we illustrate the different results we can obtain with φ_{ed} on one dataset.

D. Computational complexity of φ_{ed}

Since computing the quality of a given subgroup is a frequently occuring operation in EMM implementations, it is essential that the quality measure can be computed efficiently. In this subsection we will analyse the complexity of the calculation of φ_{ed} (and thus of φ_{weed}).

Let two Bayesian networks BN_D and BN_p be given. It is straightforward to obtain the skeletons of the networks in quadratic time. The number ℓ can also be determined in quadratic time if we have the skeletons and the moralized graphs, and after that the quality measure value is one elementary operation away. All that remains to be done is obtain the moralized graphs.

Moralized graphs can be obtained from a DAG in a straightforward manner by identifying all v-structures, drawing the resulting edges and dropping directions. The last two of these steps can be done in quadratic time, but since there can be a quadratic amount of v-structures in a DAG, and we need to visit every vertex at least once to identify them, this first step obviously costs $\mathcal{O}\left(m^3\right)$ time. Hence so does φ_{ed}.

IV. Experiments

In this section, we illustrate the usefulness of our new quality measure by finding exceptional models in several real-life datasets. We use an implementation of the EMM process that is strongly based on the Safarii Data Mining system [20]. The implementation was tailored to cope specifically with the complex target concepts needed in our EMM implementation. A run of the modified Safarii system returns all subgroups found, ranked according to φ_{weed} or φ_{ed}.

For the beam search process, we pick the following parameters. On each level, we select the $w = 10$ best subgroups, and refine these to create the candidate subgroups for the next level. This beam-width w was intentionally set to a modest value, in order to discourage too much redundancy in the reported subgroups, and to achieve a reasonable efficiency. If we want to refine a subgroup by adding a constraint on a numeric attribute, we partition that attribute into 8 equal-sized intervals, and then we consider inequalities on these dynamically allocated split points as the refining constraints. After some initial experimentation with different search depths, the maximum subgroup complexity was set to $d = 2$, i.e. a search of two levels. A larger value for d has proven to produce subgroups with irrelevant extra conditions that do not provide any benefit compared to the level 1 and 2 results. We will illustrate this effect in the presentation of the results on our first dataset.

After the main mining phase, we post-process the best t subgroups. To reduce the non-deterministic effects we outlined in Section III-B, we fit 20 Bayesian networks on the whole dataset and 20 Bayesian networks on each of the t subgroups, and assign to each subgroup as its final quality the average of the 400 resulting quality values. The effect of this post-processing is to update the quality of the t most promising subgroups to a more reliable value, and thus slightly rearranging the order of the top subgroups. For our experiments, t was set to 100, in order to guarantee that at least the top subgroups considered will be extensively evaluated. The number t can be easily set to larger values, were one so inclined.

To the best of our knowledge, validation of found patterns in subgroup discovery (hence also in EMM) is an open problem. The most common method is domain-specific interpretation of the resulting subgroups. This has the obvious drawbacks that it is subjective, and largely depends on human intuition with respect to the domain at hand: validation through this method regarding mammals that are spread over a certain geographical location will be more convincing to most humans than validation regarding probes in a phylogenetic profile. To somewhat alleviate these drawbacks, we validate each subgroup not only by domain-specific interpretation, but additionaly by showing that its exceptionality is not merely caused by random effects in the data. We do this in two ways. On the one hand, we

generate 100 random subsets of the data with the same size as the subgroup under inspection. We compute the quality of each random subset, and test whether the found subgroup quality is significantly higher than the qualities of these 100 randomly generated subsets. On the other hand, we generate 100 random patterns with the same length as the subgroup, that is, the random patterns consist of the same number of conditions as the subgroup. Furthermore, patterns were generated such that $n \geq lb_2$ (by discarding and recomputing patterns with too low a support). Then we do the same test on the quality of the found subgroup and the qualities of the subgroups corresponding to these random patterns. The idea here is that, rather than comparing our subgroup with entirely random subsets, it might be more fair to compare with random entities whose structure is similar to the structure of our subgroup.

A. Datasets

The method outlined in this paper is designed for data with multiple discrete target variables. An obvious instance of such data is multi-labeled data. Recently, Cheng and Hüllermeier published a paper [21] containing an overview of seven benchmark multi-labeled datasets, of which we selected three from different domains for our experiments.

The *emotions* dataset [22] consists of 593 songs, from which 8 rhythmic and 64 timbre features were extracted. Domain experts assigned the songs to any number of six main emotional clusters: *amazed-surprised*, *happy-pleased*, *relaxing-calm*, *quiet-still*, *sad-lonely*, and *angry-fearful*.

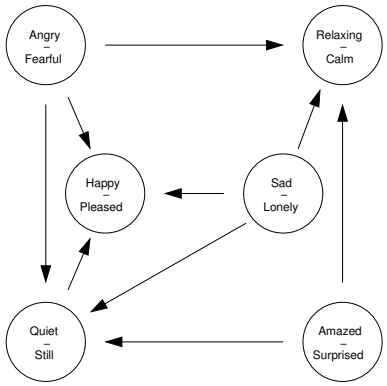
The *scene* dataset [23] is from the semantic scene classification domain, in which a photo can be classified into one or more of 6 classes. It contains 2407 photos, each of which is divided in 49 blocks using a 7×7 grid. For each block the first two spatial color moments of each band of the LUV color space are computed. This space identifies a color by its lightness (the L* band) and two chromatic valences (the u* and v* band). The photos can have the classes *beach*, *field*, *fall foliage*, *mountain*, *sunset*, and *urban*.

From the biological field we consider the *yeast* dataset [24]. It consists of micro-array expression data and phylogenetic profiles with 2417 genes of the yeast *Saccharomyces cerevisiae*. Each gene is annotated with any number of 14 functional classes.

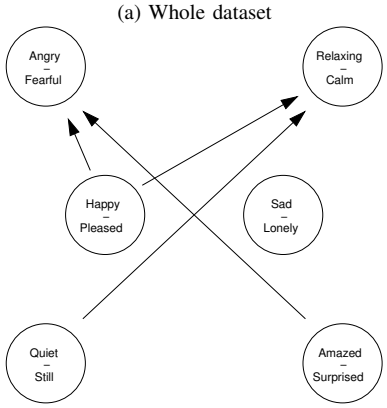
The MLC datasets all have a relatively small number of targets. Hence the fitted Bayesian networks are easy to interpret, and experiments on these datasets form a nice proof of concept for our method. However, the method is designed for larger, more complex target systems. Hence, in addition to the MLC datasets, we analyse the *mammals* dataset [25], [26]. It focuses on subdividing the geography of Europe into clusters based on their fauna, which is a core activity of biology. The dataset was created by combining two datasets: one documenting presence or absence of 101 mammals for a set of 2221 grid cells covering Europe, and one documenting climate and elevation of the corresponding land areas. We define candidate subgroups by conditions on the climate and elevation data, and fit Bayesian networks on the mammals. We use a version of this dataset that was pre-processed by Heikinheimo et al. [27].

Table I
DATASET STATISTICS

Dataset	Domain	N	k	m	Cardinality
Emotions	Music	593	72	6	1.87
Mammals	Zoogeography	2221	69	101	24.43
Scene	Vision	2407	294	6	1.07
Yeast	Biology	2417	103	14	4.24



(a) Whole dataset



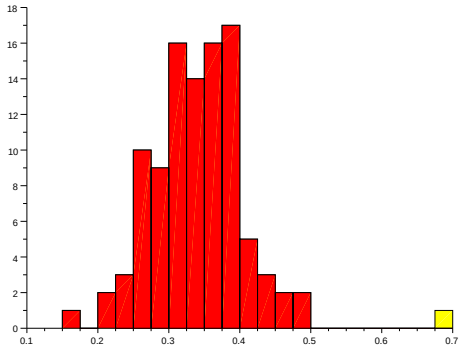
(b) *STD MFCC 7* ≤ 0.203 and *Mean Centroid* ≥ 0.066

Figure 3. Bayesian networks for the *emotions* data

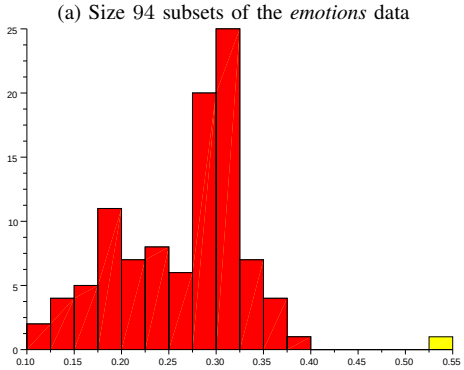
Some statistics regarding these datasets can be found in Table I. The column *Cardinality* displays the average number of positive targets per record.

B. Results

On the *emotions* dataset, we obtained the networks shown in Figure 3. Figure 3a depicts a network fitted on the whole dataset, and Figure 3b displays a network fitted on a subgroup of size 94 corresponding to the conditions *STD_MFCC_7* ≤ 0.203 and *Mean_Centroid* ≥ 0.066, with quality $\varphi_{\text{weed}} = 0.675$. The first condition says that coefficient 7 of the 13-band Mel Frequency Cepstrum has a



(a) Size 94 subsets of the *emotions* data



(b) Size 452 subsets of the *scene* data

Figure 4. Histograms for random subsets

low standard deviation, i.e. there is little variation in one of the middle spectrum bands. The second condition says that the songs in the subgroup have a moderate to high mean spectral centroid. This correlates with the impression of a bright sound [28].

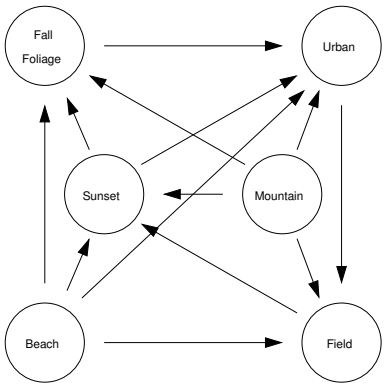
From Figure 3a we find that on the whole dataset, the emotion *sad-lonely* is correlated with all other emotions: it shares marginal dependency relations with *happy-pleased*, *relaxing-calm* and *quiet-still*, and conditional dependency relations given both *relaxing-calm* and *quiet-still* with *angry-fearful* and *amazed-surprised*. When restricted to the subgroup, *sad-lonely* is correlated with none of the other emotions (cf. Figure 3b). This seems reasonable: we would expect that bright sounds in music have a great influence on whether humans perceive a song as *sad-lonely* or not. Hence for songs with bright sounds it is more likely that *sad-lonely* is less correlated with other factors (such as the other emotions); we already have an explanation for the distribution of *sad-lonely*, so the probability that it does not depend on the other emotions increases.

The histogram of Figure 4a displays the qualities of 100 random subsets of the *emotions* data of size 94. These qualities have a mean of 0.339 and a standard deviation of 0.060. The lighter bar represents the quality of the subgroup we found, 0.675. The hypothesis that this value is generated by the normal distribution fitted to the other 100 leads to a p-value of $1.32 \cdot 10^{-8}$.

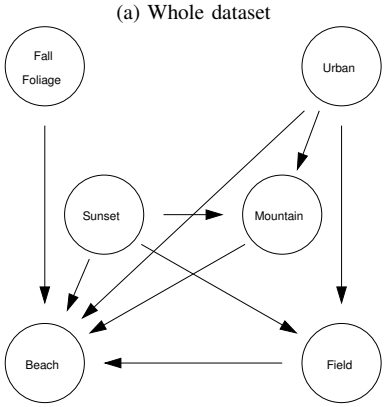
We have summarized these random benchmark results in Table II, together with the results for random patterns, and the same results for the best subgroup on each following dataset. In this table, the values for μ and σ are merely reproduced to characterize the distributions from which the relevant quantities, the p-values, are deduced.

To illustrate our choice for $d = 2$ as our maximum subgroup complexity, we also ran our algorithm on the *emotions* data with $d = 3$. The best subgroup we find is the exact same subgroup we found with $d = 2$. The next eleven subgroups in the ranking all share the two conditions that define the best subgroup. They all have one extra condition that removes at most eleven data points from the group, which leads to a slight decrease in quality. One more subgroup in the top 50 also shares these two conditions. Two other combinations of first two conditions are shared by respectively 14 and 21 subgroups in the top 50, which leaves only two subgroups that are different. These fairly homogeneous results show that running the algorithm with $d = 3$ instead of $d = 2$ is quite pointless.

Figure 5a shows the network fitted on the whole *scene* dataset. In this dataset, we found a subgroup with quality $\varphi_{\text{weed}} = 0.545$ containing 452 data points. A network fitted on the subgroup is shown in Figure 5b. The conditions

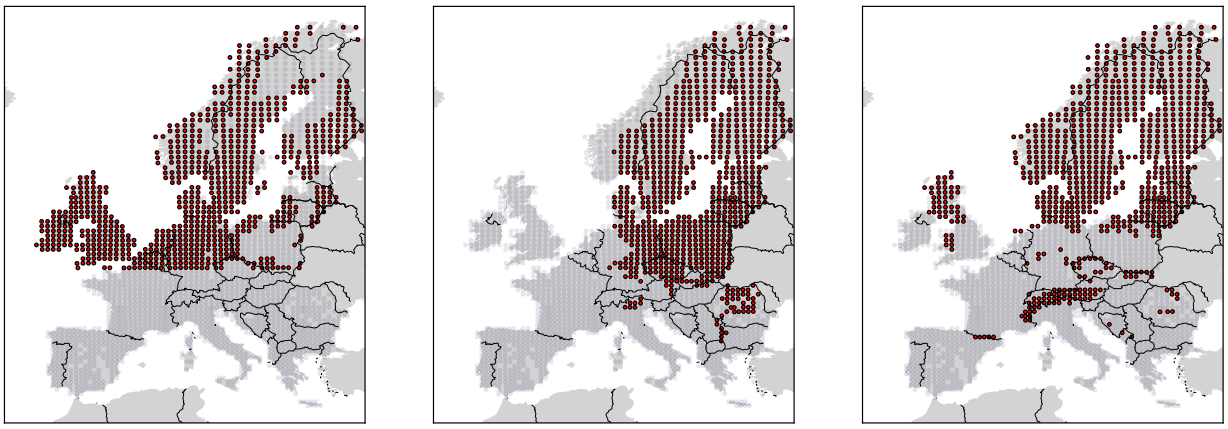


(a) Whole dataset



(b) *Mean L* band block 7* ≥ 0.699 and *Mean u* band block 19* ≤ 0.336

Figure 5. Bayesian networks for the *scene* data



(a) *latitude* ≥ 49.85 and *prec feb* ≥ 28.75 ($n = 839$)

(b) *max temp nov* ≤ 7.66 and *prec feb* ≤ 45.38 ($n = 835$)

(c) *max temp mar* ≤ 7.97 and *max temp sep* ≤ 17.65 ($n = 834$)

Figure 6. Regions in Europe that belong to the subgroups

indicate a high mean lightness in the upper right corner block of the photo, and a low mean u* chromatic valence in a more centrally located block.

Figure 4b displays the histogram of the randomly generated subsets of size 452. For a statistical validation, see again Table II.

The first-ranked subgroup on the *yeast* dataset, G_{p_1}, has quality $\varphi_{\text{weed}}(p_1) = 0.437$, and is defined by conditions on its 79-element gene expression data: *probe 3* ≤ -0.025 and *probe 66* ≥ -0.071. The next three subgroups in the ranking each share their first condition with the top-ranked subgroup, hence they are not that interesting to present here. The fifth-ranked subgroup, G_{p_5}, has quality $\varphi_{\text{weed}}(p_5) = 0.369$ and conditions *probe 9* ≤ -0.063 and *probe 53* ≥ -0.081. The subgroup sizes are $|G_{p_1}| = 681$ and $|G_{p_5}| = 530$.

From the fitted Bayesian networks, many changes in dependence relations can be deduced; we will outline a few. In G_{p_1} the functional class *cell growth, cell division, DNA synthesis* has four dependence relations less than on the whole dataset, and *protein destination* has five less. On the other hand, *energy* and *ionic homeostasis* both have an extra dependence relation. In G_{p_5}, the functional classes *cellular organization* and *cell rescue, defence, death and aging* have fewer dependence relations than on the whole dataset (six and three, respectively), while *metabolism* and *cellular biogenesis* have one more.

On the *mammals* dataset the first-ranked subgroup G_{p_1} is defined by conditions *latitude* $\geq$ 49.85 and *prec_feb* $\geq$ 28.75, i.e. northern areas with a fair amount of precipitation in February. Two other interesting subgroups (ranked sixth and eighth) are defined by meteorological conditions only. In subgroup G_{p_6} we have *max_temp_nov* $\leq$ 7.66 and *prec_feb* $\leq$ 45.38, i.e. November is not warm and precipitation in February is low, while in subgroup G_{p_8} we have *max_temp_mar* $\leq$ 7.97 and *max_temp_sep* $\leq$ 17.65, i.e. the temperatures in both March and September do not reach high levels. The subgroups have quality measure values $\varphi_{\text{weed}}(p_1) = 0.122$, $\varphi_{\text{weed}}(p_6) = 0.121 = \varphi_{\text{weed}}(p_8)$, and sizes $|G_{p_1}| = 839$, $|G_{p_6}| = 835$ and $|G_{p_8}| = 834$.

Figure 6 shows the regions in Europe that belong to the subgroups. Areas that are unique to one subgroup within this set are Ireland and the Benelux for G_{p_1} (which had the condition that it is wet in February), Romania and Poland for G_{p_6} (cold in November, dry in February), and the Alps and Pyrenees for G_{p_8} (cold in both March and September).

Among the relations between mammals that distinguish the subgroups from each other and the whole dataset D are the following: the European Water Vole (*Arvicola terrestris*) and the Mountain Hare (*Lepus timidus*) are conditionally dependent given the Ermelin (*Mustela erminea*) on D but not on any of the subgroups, only on G_{p_1} the Wildcat (*Felis silvestris*) and the Beech Marten (*Martes foina*) are conditionally dependent given the Western Roe Deer (*Capreolus capreolus*), only on G_{p_6} the Broad-toothed Field Mouse (*Apodemus mysticanus*) and the Lesser Mole Rat (*Nannospalax leucodon*) are conditionally dependent given the Marbled Polecat (*Vormela peregusna*), and only on G_{p_8} the Red Squirrel (*Sciurus vulgaris*) and the Least Weasel (*Mustela nivalis*) are conditionally dependent given the European Badger (*Meles meles*).

In Section III-C we claimed that we needed the entropy term in our quality measure to avoid obtaining small subgroups. To substantiate that claim, we also ran our algorithm on the *mammals* dataset with φ_{ed} instead of φ_{weed}. The first ranked subgroup we found with this distance has size 105 and is defined by conditions *mean_temp_apr* $\geq$ 11.86 and *mean_temp_aug* $\leq$ 23.28. The regions in Europe that belong to this subgroup are displayed in Figure 7. Notice that although this group is smaller than those found with φ_{weed}, it may still be interesting.

Table II
RANDOM BENCHMARK RESULTS OF BEST SUBGROUPS

Dataset	$\varphi_{\text{weed}}(G_{p_1})$	Random subsets			Random patterns		
		μ	σ	p-value	μ	σ	p-value
Emotions	0.675	0.339	0.060	$1.32 \cdot 10^{-8}$	0.302	0.114	$5.48 \cdot 10^{-4}$
Scene	0.545	0.263	0.065	$6.07 \cdot 10^{-6}$	0.319	0.085	0.004
Yeast	0.437	0.296	0.032	$5.57 \cdot 10^{-6}$	0.250	0.046	$2.19 \cdot 10^{-5}$
Mammals	0.122	0.072	0.005	$1.43 \cdot 10^{-21}$	0.094	0.017	0.043
Mammals (φ_{ed})	0.147	0.125	0.007	0.002	0.107	0.014	0.002

Considering again Table II, we note that the reported subgroups have a significantly higher quality than randomly generated subsets with the same size, and to a lesser extent also a significantly higher quality than subsets corresponding to random patterns. The positive comparison to random subsets tells us that descriptive information in the attributes a_i concerning dependencies between targets is being exploited, and furthermore, useful conditions are being found by the search algorithm. The statistical validation using random patterns is a more strict validation, as it eliminates the factor of exploitation of $\overrightarrow{a}$, and primarily validates whether the measure-guided beam search is able to effectively select high-quality subgroups. As such, the somewhat lower significance levels are to be expected.

V. CONCLUSIONS AND FURTHER RESEARCH

We propose to use the interdependencies between discrete target variables as an exceptionality measure for subgroups. These interdependencies are modeled by Bayesian networks, and the quality of a subgroup is defined as the difference between the network on the whole dataset and the network on the subgroup. To quantify this difference and thus the exceptionality of the model, we define a distance metric on Bayesian networks with the same vertex set. As a post-processing step, the impact of the non-determinism in the induction of Bayesian networks is reduced by repeated modeling. Finally, statistical validation using both random subsets and random patterns demonstrates that significant findings in four domains can be made.

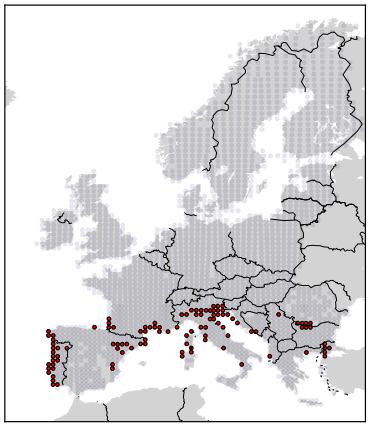


Figure 7. Regions in Europe that belong to the subgroup corresponding to *mean temp apr* ≥ 11.86 and *mean temp aug* ≤ 23.28 ($n = 105$)

The work presented in this paper can be extended in various ways. For instance, we could integrate our approach with the approach presented by Leman et al., who determined the exceptionality of a subgroup by comparing underlying probability distributions using Hellinger distance [1], [29]. Considering the Bayesian network parameters, or merely the signs of the correlations for ordered variables, could also improve our method. Furthermore, if a faster way to find the essential graph of a Bayesian network is found, we can employ it to improve the quality measure we defined.

Also, the computational burden of the edit distance for Bayesian networks may be alleviated. In Section III-D we described how we can compute the distance in quadratic time except for the bottleneck that costs $\mathcal{O}\left(m^3\right)$ time: determine all v-structures. However, this can be done by matrix multiplication: detect the v-structures in a graph by multiplying its incidence matrix with the incidence matrix of the graph with all directions reversed. Hence, we could improve our algorithm to $\mathcal{O}\left(m^{2.376}\right)$ time by using the Coppersmith-Winograd algorithm [30]. Unfortunately, the implied constant term is so large that the algorithm becomes impractical. Still, the use of other algorithms such as Strassen's [31], which takes $\mathcal{O}\left(m^{2.807}\right)$, might be beneficial.

Perhaps the most promising direction in which this EMM approach could be employed would be its use in the Local Pattern Discovery phase in the LeGo framework [32]. As our subgroups identify parts of the input space where exceptional sets of dependencies hold, they can be thought of as a means to simplify a given multi-label classification problem, by allowing for different classification models in different subgroups. As subgroups may represent more coherent samples of the data, compared to the whole database, it can be expected that the LeGo building blocks can be employed to improve predictive accuracy.

ACKNOWLEDGMENTS

This research is financially supported by the Netherlands Organisation for Scientific Research (NWO) under project number 612.065.822. The European mammals data was kindly provided by Tony Mitchell-Jones and the Societas Europaea Mammalogica.

References

[1] D. Leman, A. Feelders, A. J. Knobbe, Exceptional Model Mining, Proc. ECML/PKDD (2) 2008, LNCS, volume 5212, pp. 1–16, Springer, Heidelberg.

[2] L. G. Shapiro, R. M. Haralick, A Metric for Comparing Relational Descriptions, IEEE Trans. Pattern Anal. Mach. Intell. 7, pp. 90–94, 1985.

[3] A. K. McCallum, Multi-label text classification with a mixture model trained by EM, AAAI 1999 Workshop on Text Learning.

[4] R. T. Paine, Food Web Complexity and Species Diversity, The American Naturalist 100 (910), pp. 65–75, 1966.

[5] G. A. Davis, Bayesian reconstruction of traffic accidents, Law, Probability and Risk 2, pp. 69–89, 2003.

[6] F. J. Díez, J. Mira, E. Iturralde, S. Zubillaga, DIAVAL, a Bayesian expert system for echocardiography, Artificial Intelligence in Medicine 10, pp. 59–73, 1997.

[7] M. Neil, N. Fenton, M. Tailor, Using Bayesian Networks to Model Expected and Unexpected Operational Losses, Risk Analysis 25 (4), 2005.

[8] Y. H. Xu, A. Fern, On Learning Linear Ranking Functions for Beam Search, Proc. ICML 2007, ACM International Conference Proceeding Series vol. 227, pp. 1047–1054, ACM, New York.

[9] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, San Francisco, 1988.

[10] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, D. J. Spiegelhalter, Probabilistic Networks and Expert Systems, Springer-Verlag, New York, pp. 31–33, 1999.

[11] W. L. Buntine, Theory Refinement on Bayesian Networks, Proc. UAI 1991, pp. 52–60, Morgan Kaufmann, San Francisco.

[12] D. Heckerman, A Tutorial on Learning with Bayesian Networks, Proc. NATO Advanced Study Institute on Learning in Graphical Models, pp. 301–354, Kluwer Academic Publishers, Norwell, MA, 1998.

[13] P. Larrañaga, M. Poza, Y. Yurramendi, R. H. Murga, C. M. H. Kuijpers, Structure Learning of Bayesian Networks by Genetic Algorithms: A Performance Analysis of Control Parameters, IEEE Trans. Pattern Anal. Mach. Intell. 18 (9), pp. 912–926, 1996.

[14] D. Heckerman, D. Geiger, D. M. Chickering, Learning Bayesian networks: The combination of knowledge and statistical data, Machine Learning 20, pp. 197–243, 1995.

[15] D. M. Chickering, A Transformational Characterization of Equivalent Bayesian Network Structures, Proc. UAI 1995, pp. 87–98, Morgan Kaufmann, San Francisco.

[16] C. Riggelsen, Approximation Methods for Efficient Learning of Bayesian Networks, IOS Press, Amsterdam, 2008.

[17] S. A. Andersson, D. Madigan, M. D. Perlman, A Characterization of Markov Equivalence Classes for Acyclic Digraphs, Annals of Statistics, 25 (2), pp. 505–541, 1997.

[18] D. M. Chickering, Learning Equivalence Classes of Bayesian-Network Structures, Journal of Machine Learning Research 2, pp. 445–498, 2002.

[19] T. Verma, J. Pearl, Equivalence and Synthesis of Causal Models, Proc. UAI 1990, pp. 255–270, Elsevier, Amsterdam.

[20] A. J. Knobbe, Safarii multi-relational data mining environment, 2006, http://www.kiminkii.com/safarii.html

[21] W. Cheng, E. Hüllermeier, Combining Instance-Based Learning and Logistic Regression for Multilabel Classification, Proc. ECML/PKDD (1) 2009, LNCS, volume 5781, pp. 6, Springer, Heidelberg.

[22] K. Trohidis, G. Tsoumakas, G. Kalliris, I. P. Vlahavas, Multi-Label Classification of Music into Emotions, Proc. 9th International Conference on Music Information Retrieval, pp. 325-330, 2008.

[23] M. R. Boutell, J. Luo, X. Shen, C. M. Brown, Learning Multi-Label Scene Classification, Pattern Recognition 37 (9), pp. 1757-1771, 2004.

[24] A. Elisseeff, J. Weston, A Kernel Method for Multi-Labelled Classification, Advances in Neural Information Processing Systems 14, pp. 681–687, MIT Press, Cambridge MA, 2002.

[25] G. C. Garriga, H. Heikinheimo, J. K. Seppänen, Cross-mining binary and numerical attributes, Proc. ICDM 2007, pp. 481-486.

[26] T. Mitchell-Jones et al.: The Atlas of European Mammals, Poyser natural history, 1999.

[27] H. Heikinheimo, M. Fortelius, J. Eronen, H. Manilla, Biogeography of European land mammals shows environmentally distinct and spatially coherent clusters, Journal of Biogeography 34 (6), pp. 1053–1064, 2007.

[28] E. Schubert, J. Wolfe, A. Tarnopolsky, Spectral centroid and timbre in complex, multiple instrumental textures, Proc. 8th International Conference on Music Perception & Cognition, pp. 654–657, 2004.

[29] G. Yang, L. Le Cam, Asymptotics in Statistics: Some Basic Concepts, Springer, Berlin, 2000.

[30] D. Coppersmith, S. Winograd, Matrix multiplication via arithmetic progressions, Journal of Symbolic Computation 9 (3), pp. 251–280, 1990.

[31] V. Strassen, Gaussian Elimination is not Optimal, Numerische Mathematik 13, pp. 354–356, 1969.

[32] A. J. Knobbe, B. Crémilleux, J. Fürnkranz, M. Scholz, From Local Patterns to Global Models: The LeGo Approach to Data Mining, Proc. ECML/PKDD 2008 LeGo workshop, pp. 1–16.