Exceptional Model Mining

Duivesteijn, W.

Chapter 6

Unusual Conditional Interactions – Bayesian Network Model

In Chapter 4, we discussed an EMM instance with an internally unsupervised model class, regarding the correlation between two attributes. In Chapter 5, we discussed an EMM instance with an internally supervised model class, classifying a single output target attribute based on one or several input target attributes. Depending on the choice of classifier, this may or may not incorporate complex interactions between sets of input target attributes; in any case, such complex interactions have not yet been considered for an unsupervised model class. In this chapter we fill that void, by considering the Exceptional Model Mining instance with a Bayesian network as model class.

In the Bayesian network model class we allow multiple nominal targets $\ell_1, \ldots, \ell_m$. A description is deemed interesting, when the conditional dependence relations between the targets are substantially different for the description from these relations on the whole dataset. Hence we validate the descriptions on the conditional interdependencies between the targets, rather than the target values themselves. To capture these interdependencies, we learn a Bayesian network between the targets, from data.

The choice to capture complex interactions between larger sets of unsupervised target attributes by means of conditional dependence relations, is inspired by the *Pisaster* example discussed in Chapter 2. Recall that the field study of Robert T. Paine [86] yields, among many other results, that a

conditional dependence relation exists between the sponge *Haliclona*, the nudibranch *Anisodoris*, and the starfish *Pisaster ochraceus*. This study gives a real-life example of multiple-target interactions that require the complexity of a Bayesian network.

There are many algorithms to learn a Directed Acyclic Graph (DAG) model, such as a Bayesian network, from data; see for instance [8, 47, 67]. 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 Section 5.3.1), 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 [12], in order to escape from a maximum that may be local. For more details on this combination of methods, see [95].

Notice that this process is 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 find different Bayesian networks for the same data with different random seeds. However, these variations were modest: few edges change, and resulting networks for the same data are usually equivalent.

We consider the choice of method to learn a Bayesian network from data a parameter of this EMM instance.

6.1 Quality Measure φ_{weed}

Having chosen a method to learn a Bayesian network from data, we would like to employ such networks to capture deviating conditional dependence relations between targets. Our quality measure uses the structure of the learned networks to this end. The main idea is to start the EMM process by learning a Bayesian network BN^{Ω} between the targets from the entire

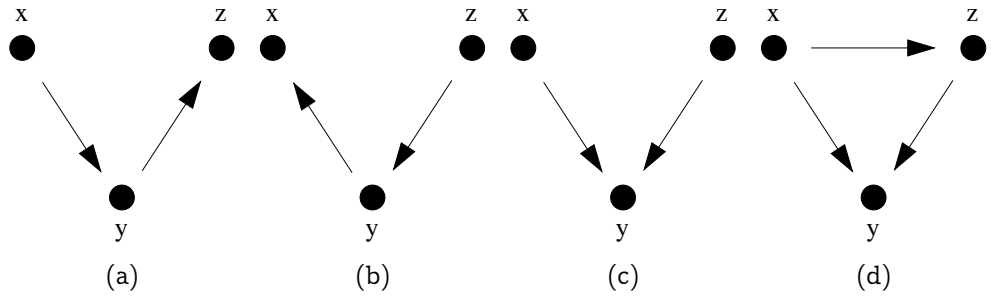


Figure 6.1: Example Bayesian networks.

dataset. Then, for each description D under consideration, we learn another Bayesian network BN^D, but we learn it *only from the records covered by* D. Comparing the structure of the networks BN^Ω and BN^D then gives us a measure for the quality of the description D. One might be tempted to consider traditional edit distance between graphs to make this comparison, but then we would not take into account some peculiarities about how Bayesian networks represent independence relations.

6.1.1 Independence Relations in Bayesian Networks

There are two important peculiarities about the independence relations in Bayesian networks, which we illustrate by the example networks in Figure 6.1. First, seemingly 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), 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 chapter. It is a structure as seen in network (c): a *v-structure*.

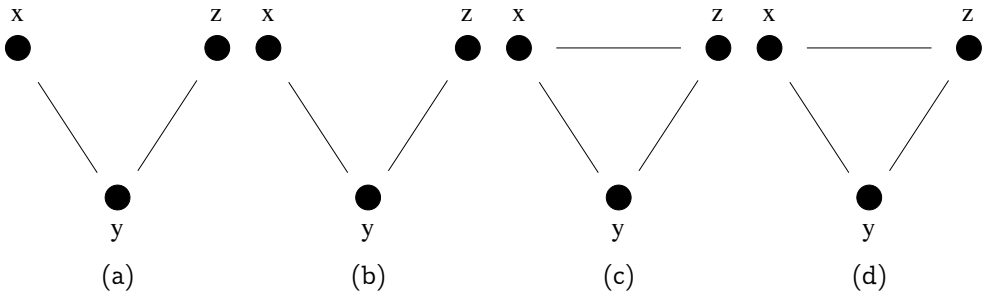


Figure 6.2: Moralized graphs for the networks in Figure 6.1.

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. A v-structure is also known as an *immorality*, since the parents of vertex y are 'unmarried', i.e. there is no edge between them. A graph can be *moralized* [17] 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 6.1 are depicted in Figure 6.2. As one can see, the moralized version of network (c) has an extra edge, which corresponds to removing the v-structure from the original network.

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

6.1.2 Edit Distance for Bayesian Networks

To overcome the peculiarities of Bayesian networks, we propose a heuristic quality measure based on the following well-known result by Verma and Pearl [111]

Theorem 2 (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 potential edges violating the conditions 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. 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

$$\zeta = \left| \left[S^1 \ominus S^2\right] \cup \left[M^1 \ominus M^2\right] \right|$$

The distance between BN^1 and BN^2 is defined as

$$\delta(BN^1, BN^2) = \frac{2\zeta}{m(m-1)}$$

As usual in set theory, $\ominus$ denotes a symmetric difference: $X \ominus Y = (X \cup Y) - (X \cap Y)$. 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 targets in the dataset, hence vertices in the Bayesian networks.

We illustrate the edit distance by computing the mutual distances between the networks in Figure 6.1. We find that $\delta(a, b) = 0$ and $\delta(a, c) = \delta(a, d) = \delta(b, c) = \delta(b, d) = \delta(c, d) = 1/3$. Only for the two networks that are equivalent, distance 0 is obtained. If we compare the networks to the independence model $\varnothing$ which has no edges at all, we obtain $\delta(a, \varnothing) = \delta(b, \varnothing) = 2/3$, and $\delta(c, \varnothing) = \delta(d, \varnothing) = 1$.

The edit distance can now be used to quantify the exceptionality of a description

Definition (Edit distance based quality measure). Let a description D be given. Denote the Bayesian network we learn from Ω by BN^{Ω}, and denote the Bayesian network we learn from G_D by BN^D. Then the quality of D is

$$\varphi_{ed}(D) = \delta\left(BN^{\Omega}, BN^D\right)$$

If we would plug φ_{ed} into the EMM framework, a familiar problem 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 combine the measure with the entropy function φ_{ef} (cf. Section 3.2), to obtain the following aggregate measure.

Definition (Weighed Entropy and Edit Distance).

$$\varphi_{\text{weed}}(D) = \sqrt{\varphi_{\text{ef}}(D)} \cdot \varphi_{\text{ed}}(D)$$

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 description with large edit distance, while mediocre entropy is acceptable.

6.2 Experiments

6.2.1 Datasets

The *Emotions* dataset [103] 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 from the Tellegen-Watson-Clark model of mood [102]: *'amazed-surprised'*, *'happy-pleased'*, *'relaxing-calm'*, *'quiet-still'*, *'sad-lonely'*, and *'angry-fearful'*.

The *Scene* dataset [6] 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 into 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 [28]. 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.

Table 6.1: Statistics concerning the datasets used in the Bayesian Network Model and Multi-label LeGo experiments (cf. Chapter 9). Here, N is the total number of records, k is the number of descriptive attributes, and m is the number of nodes in the fitted Bayesian network model. The column *Cardinality* displays the average number of positive targets per record.

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

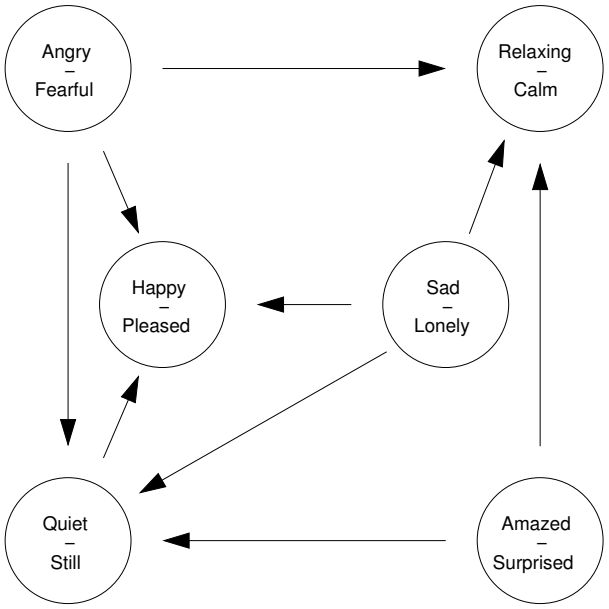
The three introduced datasets all have a relatively small number of targets. Hence the fitted Bayesian networks are easy to interpret, and experiments on these datasets form a nice proof of concept for our method. However, EMM with the Bayesian Network model class can also handle larger, more complex target systems. Hence, in addition to the MLC datasets, we analyse the *Mammals* dataset [40, 80]. 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. [49].

Some statistics regarding these datasets can be found in Table 6.1.

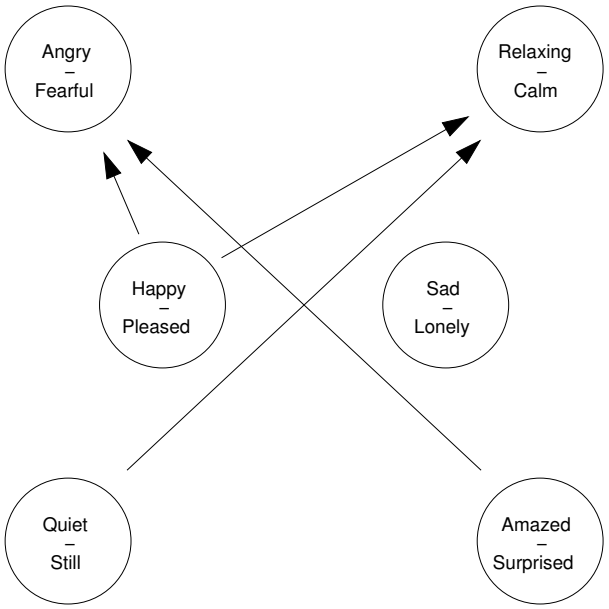
6.2.2 Experimental Results

Emotions Data

On the *Emotions* dataset, we obtained the networks shown in Figure 6.3. Figure 6.3a depicts a network learned from the whole dataset, and Figure 6.3b displays a network learned from a subgroup of size 94 (15.9%) corresponding to description $D_6 : STD\_MFCC\_7 \leq 0.203 \wedge Mean\_Centroid \geq 0.066$, with quality $\varphi_{\text{weed}}(D_6) = 0.675$. The first condition says that coef-



(a) Whole dataset.



(b) D_6 : $STD\_MFCC\_7 \leq 0.203 \wedge Mean\_Centroid \geq 0.066$.

Figure 6.3: Bayesian networks for the *Emotions* data.

ficient 7 of the 13-band Mel Frequency Cepstrum has a low standard deviation, which has a nontrivial interpretation. 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 [96].

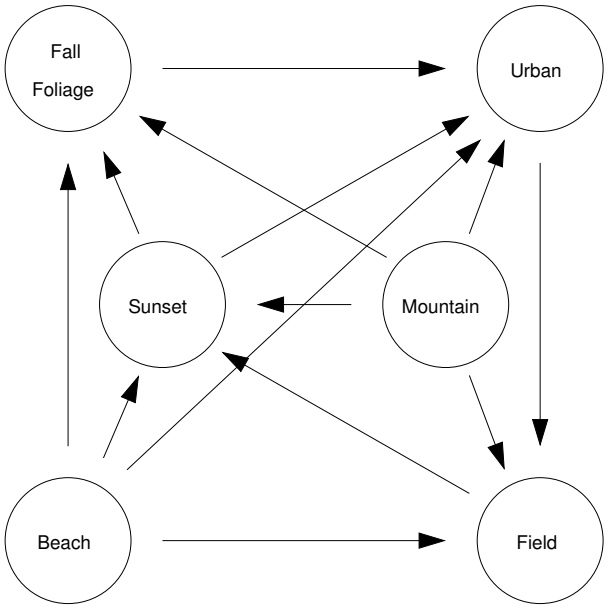
From Figure 6.3a we find that on the whole dataset, the emotion *sad-lonely* is correlated with all other emotions: it shares marginal dependence relations with *happy-pleased*, *relaxing-calm* and *quiet-still*, and conditional dependence relations given both *relaxing-calm* and *quiet-still* with *angry-fearful* and *amazed-surprised*. When restricted to the description, *sad-lonely* is correlated with none of the other emotions (cf. Figure 6.3b). 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 increases that it does not depend on the other emotions.

Scene Data

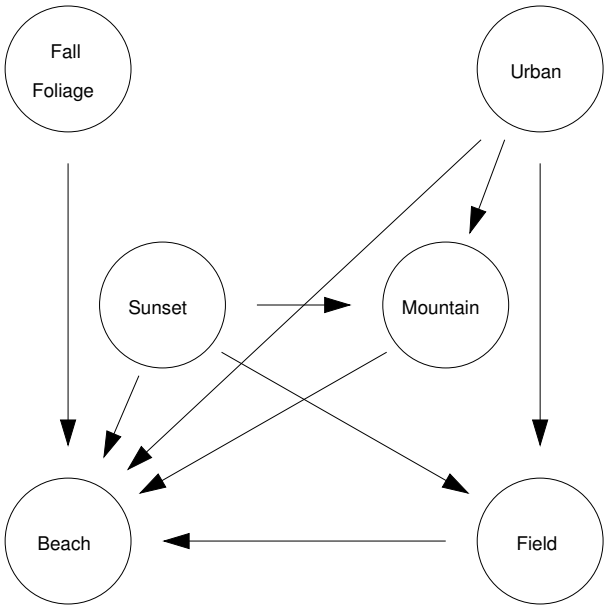
Figure 6.4 shows the networks fitted on the *Scene* dataset. In this dataset, we found a description with quality $\varphi_{\text{weed}}(D_7) = 0.545$, covering 452 records (18.8%). The conditions indicate a high mean lightness in the upper right corner of the photo, and a low mean u* chromatic valence in a more centrally located area.

Yeast Data

The first-ranked description on the *Yeast* dataset has quality $\varphi_{\text{weed}}(D_8) = 0.437$, and is defined by conditions on its 79-element gene expression data: *probe* $3 \leq -0.025 \wedge$ *probe* $66 \geq -0.071$. The three subsequent descriptions in the ranking each share their first condition with the top-ranked descriptions, hence they are not that interesting to present here. The fifth-ranked description has quality $\varphi_{\text{weed}}(D_9) = 0.369$ and conditions *probe* $9 \leq -0.063 \wedge$ *probe* $53 \geq -0.081$. The subgroup sizes are $|G_8| = 681$ (28.2%) and $|G_9| = 530$ (21.9%).



(a) Whole dataset.



(b) D_7 : *Mean L* band block 7* $\geq 0.699 \wedge$ *Mean u* band block 19* ≤ 0.336.

Figure 6.4: Bayesian networks for the *Scene* data.

From the fitted Bayesian networks, many changes in dependence relations can be deduced; we will outline a few. In G_8 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_9, 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.

Mammals Data

On the *Mammals* dataset, the first-ranked description D_{10} is defined by conditions $latitude \geq 49.85 \wedge prec\_feb \geq 28.75$, i.e. northern areas with a fair amount of precipitation in February. Two other interesting descriptions (ranked sixth and eighth) are defined by meteorological conditions only. In description D_{11} we have $max\_temp\_nov \leq 7.66 \wedge prec\_feb \leq 45.38$, i.e. November is not warm and precipitation in February is low, while in description D_{12} we have $max\_temp\_mar \leq 7.97 \wedge max\_temp\_sep \leq 17.65$, i.e. the temperatures in both March and September do not reach high levels. The descriptions have quality $\varphi_{\text{weed}}(D_{10}) = 0.122$, $\varphi_{\text{weed}}(D_{11}) = 0.121 = \varphi_{\text{weed}}(D_{12})$, and coverage $|G_{10}| = 839$ (37.8%), $|G_{11}| = 835$ (37.6%), and $|G_{12}| = 834$ (37.6%).

The Figures 6.5, 6.6, and 6.7 chart the regions in Europe that belong to the descriptions. Areas that are unique to one description within this set are Ireland and the Benelux for D_{10} (which had the condition that it is wet in February), Romania and Poland for D_{11} (cold in November, dry in February), and the Alps and Pyrenees for D_{12} (cold in both March and September).

Among the relations between mammals that distinguish the descriptions from each other and the whole dataset Ω are the following: the European Water Vole (*Arvicola terrestris*) and the Mountain Hare (*Lepus timidus*) are conditionally dependent given the Ermelin (*Mustela erminea*) on Ω but not on any of the descriptions, only on D_{10} the Wildcat (*Felis silvestris*) and the Beech Marten (*Martes foina*) are conditionally depen-

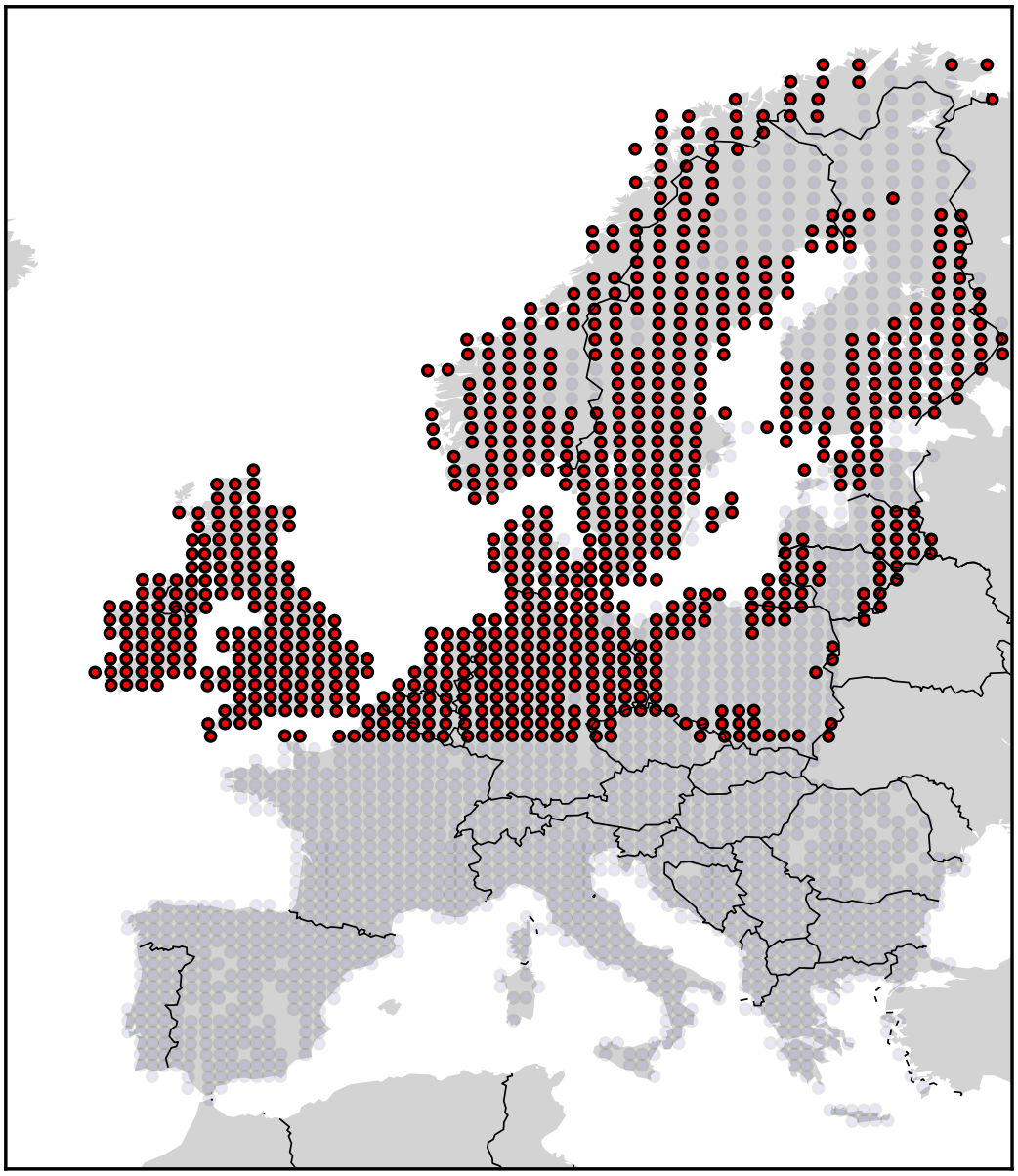


Figure 6.5: Regions in Europe that belong to the subgroup corresponding to D_{10} : $latitude \geq 49.85 \wedge prec\_feb \geq 28.75$ ($|G_{10}| = 839$).

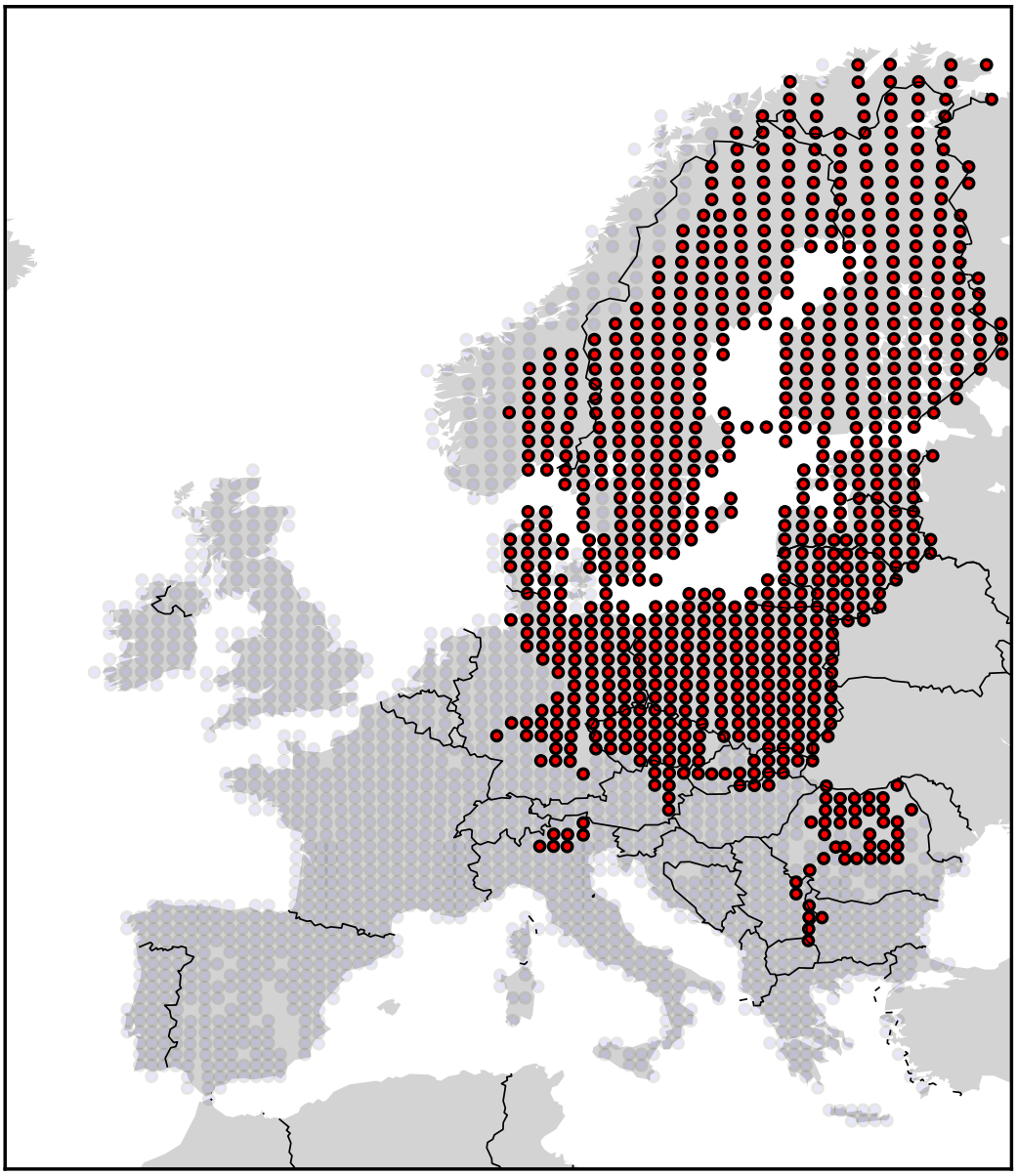


Figure 6.6: Regions in Europe that belong to the subgroup corresponding to D_{11} : $max\_temp\_nov \leq 7.66 \wedge prec\_feb \leq 45.38$ ($|G_{11}| = 835$).

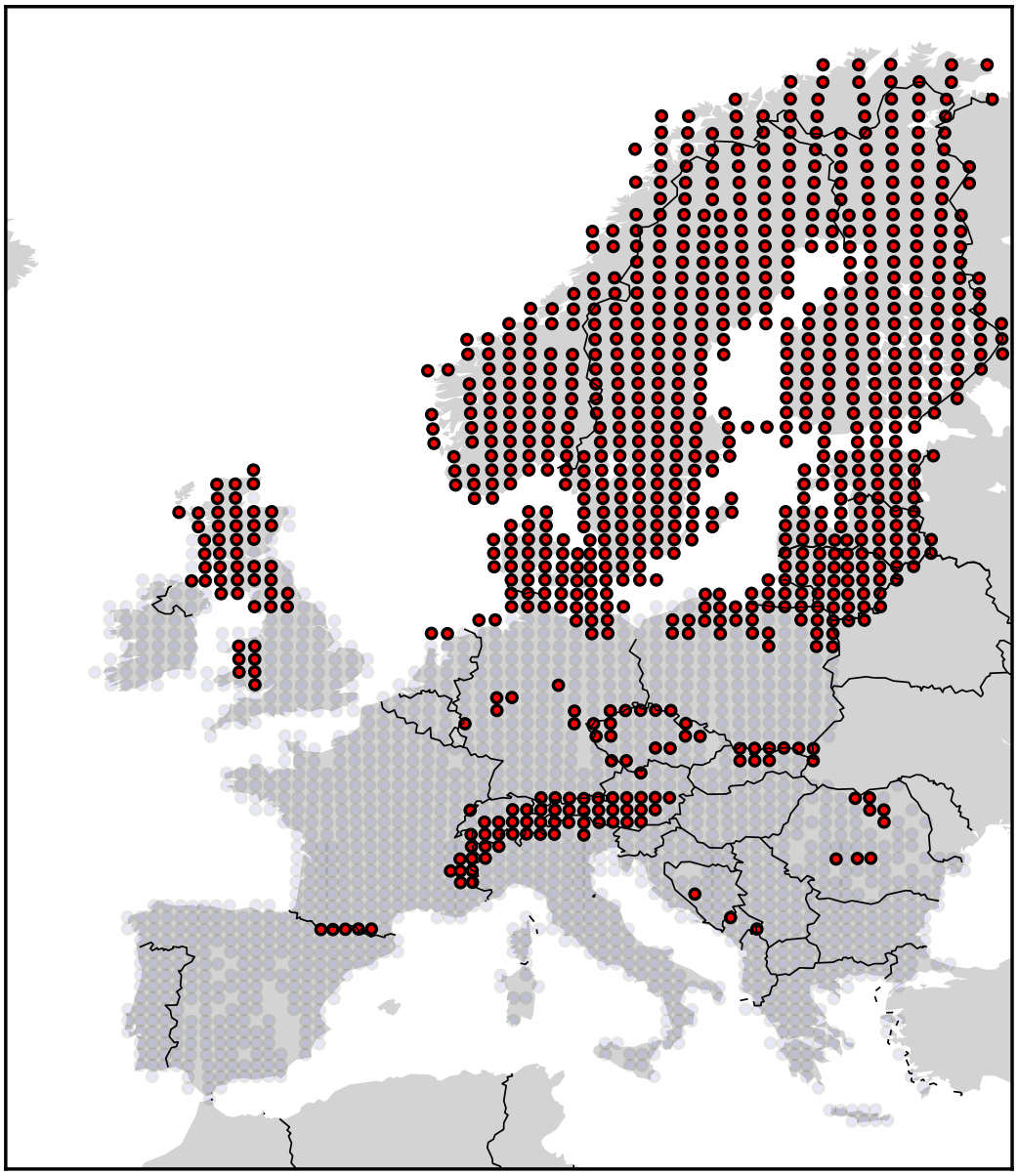


Figure 6.7: Regions in Europe that belong to the subgroup corresponding to $D_{12} : max\_temp\_mar \leq 7.97 \wedge max\_temp\_sep \leq 17.65$ ($|G_{12}| = 834$).

dent given the Western Roe Deer (*Capreolus capreolus*), only on D_{11} 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 D_{12} the Red Squirrel (*Sciurus vulgaris*) and the Least Weasel (*Mustela nivalis*) are conditionally dependent given the European Badger (*Meles meles*).

6.3 Alternatives

In Section 6.1.2, we discussed how we incorporated an entropy term in our quality measure φ_{weed}, in order to avoid obtaining small subgroups. If small subgroups are required, we can also run this EMM instance with the non-composite quality measure φ_{ed}, selecting the good descriptions only by virtue of their edit distance on Bayesian networks. To illustrate what the outcome of such a run can be, we repeated the experiments from the previous section on the *Mammals* dataset with φ_{ed} instead of φ_{weed}. The first-ranked description we found with this distance is D_{13} : *mean_temp_apr* $\geq 11.86 \wedge$ *mean_temp_aug* ≤ 23.28. Its quality is $\varphi_{\text{ed}}(D_{13}) = 0.147$, and its coverage is $|G_{13}| = 105$ (4.7%). The regions in Europe that belong to this description are displayed in Figure 6.8.

The relations between mammals that distinguish D_{13} from Ω include the following. On Ω, but not on D_{13}, the Alpine Marmot (*Marmota marmota*) and the Alpine Field Mouse (*Apodemus alpicola*) are conditionally dependent given the Alpine Ibex (*Capra ibex*), and the Beech Marten (*Martes foina*) and the Red Fox (*Vulpes vulpes*) are conditionally dependent given the Least Weasel (*Mustela nivalis*). On D_{13}, but not on Ω, the Common Genet (*Genetta genetta*) and the European Mink (*Mustela lutreola*) are conditionally dependent given the Crowned Shrew (*Sorex coronatus*), and the European Snow Vole (*Chionomys nivalis*) and the Iberian Shrew (*Sorex granarius*) are conditionally dependent given the Lusitanian Pine Vole (*Microtus lusitanicus*).

Using plain φ_{ed} instead of the composite φ_{weed} has its benefits and its drawbacks. When we compare the description D_{13} found with φ_{ed}, with the descriptions D_{10}, D_{11}, and D_{12} found with φ_{weed}, there are several things to remark. As expected, using the plain edit distance leads EMM to report

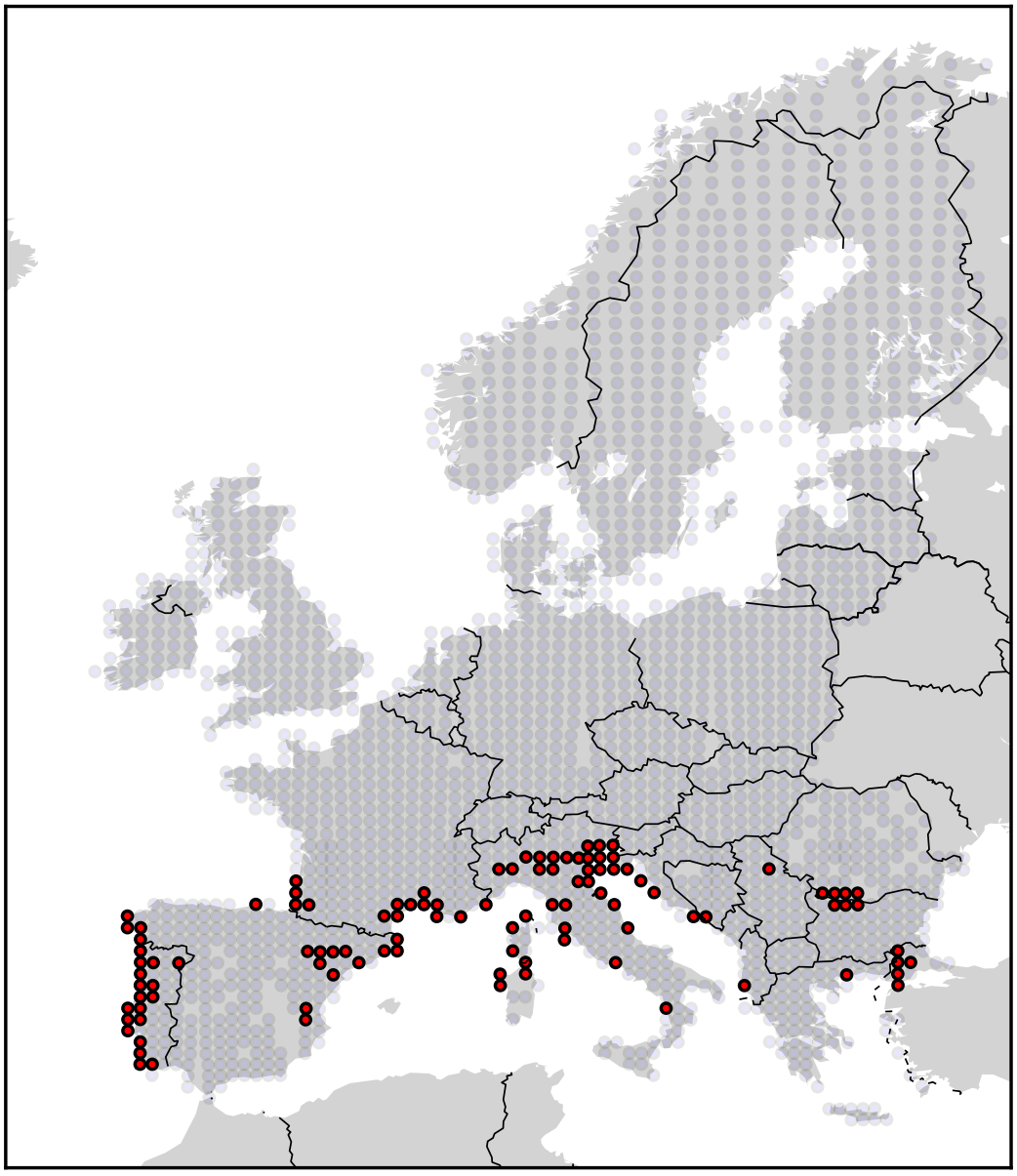


Figure 6.8: Regions in Europe that belong to the subgroup corresponding to D_{13} : $mean\_temp\_apr \geq 11.86 \wedge mean\_temp\_aug \leq 23.28$ ($|G_{13}| = 105$).

smaller subgroups than we obtain when using the edit distance weighted with entropy. Whether this is an argument for using φ_{ed} or φ_{weed} depends on the problem statement or domain expert at hand.

When we look at the deviating conditional dependence relations between the mammals, we find that particularly in the description found with the plain edit distance, the relations tend to focus on mammals that appear only in a very small subarea of Europe. For instance, within the parts of Europe covered by the dataset, the European Mink only occurs in a small area in the South West of France and the North of Spain, while the Iberian Shrew and the Lusitanian Pine Vole are confined to the Iberian peninsula. So, roughly speaking, φ_{ed} can be seen as more focused than φ_{weed}.

On the other hand, if we look at the maps of regions of Europe belonging to the subgroups, we see that φ_{weed} finds subgroups that are, geographically speaking, more coherent than the subgroup found with φ_{ed}. As we can see in Figure 6.5, subgroup G_{10} spans the North West of Europe, and as we can see in Figure 6.6, subgroup G_{11} spans the North East of Europe. At first glance, the area depicted in Figure 6.7 seems to indicate that subgroups G_{12} spans a dichotomous part of Europe: part is coherent, spanning Scandinavia, Scotland, Wales, and the Baltic countries, but to the South of that we find what appears to be rubble. However, if we compare this chart to a map of Europe indicating altitude, we find that the "rubble" actually largely overlaps with mountainous areas: we have found the Alps, the Pyrenees, the Harz, and the Carpathians. So, G_{12} spans some Northern areas, and some mountainous areas. By contrast, the regions belonging to subgroup G_{13}, as depicted in Figure 6.8, are far more scattershot. The coastal line of Portugal is a fairly coherent part of the subgroup, but the remaining areas seem relatively random. Although "mediterranean coastal" is a recurring theme, the selection of parts of the mediterranean coast seems incoherent, as does the isolated grid cell in Serbia and the small chunks in Bulgaria and Turkey. Hence, roughly speaking, φ_{weed} seems to deliver more substantially coherent subgroups than φ_{ed}.

6.4 Conclusions

In this chapter, we propose to use the interdependencies between discrete target variables as an exceptionality measure for descriptions. These interdependencies are modeled by Bayesian networks, and the quality of a description 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. Experiments show that substantial findings on four domains can be made.

Compared to the previous two chapters, the model class in the current chapter is substantially more complex. This allows EMM to search for deviations in sophisticated interplay between multiple targets simultaneously. However, the price we pay for this advantage, is that interpreting results becomes problematic. As always, the found descriptions themselves can still be interpreted easily by a domain expert. Whether interpretation of the associated models is possible, however, depends on the number of targets in the dataset at hand.

As we have seen in our analysis of the results on the *Emotions* and *Scene* datasets, we can obtain meaningful insights from comparing Bayesian networks having six vertices. However, on the *Yeast* dataset the Bayesian network contains fourteen vertices, and on the *Mammals* dataset the network contains 101 vertices. For such large networks, we can still analyze the models associated with descriptions in a limited way, by highlighting dependence relations in small subsets of the vertices that differ between the description and the whole dataset. Having an overview of deviating (conditional) dependence relations between entire networks, however, has become impossible.

In such cases, it helps when the dataset has a third set of attributes, in addition to the descriptors and the targets. In the *Mammals* dataset, such a third set is available: the location information of grid cells throughout Europe. If a description, defined on the first set of attributes and evaluated on the second set, also displays coherence on the third set of attributes, then this reinforces our belief that we have found something substantial in our dataset. For instance, the fact that the geographically coherent region of

the Alps is highlighted in Figure 6.7, even though D_{12} was neither defined nor evaluated on location information, is strong corroborating evidence that this description indicates an actual underlying phenomenon in the dataset.

The work presented in this chapter can be extended in various ways. For instance, we could integrate our approach with the Hellinger distance introduced in Section 5.3.2, to determine the exceptionality of a description by comparing underlying probability distributions. Considering the Bayesian network parameters, or merely the signs of the correlations for ordered variables, could also improve our method.

Perhaps the most promising direction in which this EMM approach could be employed will be explored in Chapter 9: as a building block to be used in the Local Pattern Discovery phase in the LeGo framework [57]. As our descriptions 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 descriptions. As descriptions 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

The European mammals data was kindly provided by Tony Mitchell-Jones and the Societas Europaea Mammalogica.