Data Min Knowl Disc (2016) 30:47–98
DOI 10.1007/s10618-015-0403-4

Exceptional Model Mining

Supervised descriptive local pattern mining with complex target concepts

Wouter Duivesteijn · Ad J. Feelders · Arno Knobbe

Received: 9 August 2013 / Accepted: 22 January 2015 / Published online: 4 February 2015

Abstract Finding subsets of a dataset that somehow deviate from the norm, i.e. where something interesting is going on, is a classical Data Mining task. In traditional local pattern mining methods, such deviations are measured in terms of a relatively high occurrence (frequent itemset mining), or an unusual distribution for one designated target attribute (common use of subgroup discovery). These, however, do not encompass all forms of "interesting". To capture a more general notion of interestingness in subsets of a dataset, we develop Exceptional Model Mining (EMM). This is a supervised local pattern mining framework, where several target attributes are selected, and a model over these targets is chosen to be the target concept. Then, we strive to find subgroups: subsets of the dataset that can be described by a few conditions on single attributes. Such subgroups are deemed interesting when the model over the targets on the subgroup is substantially different from the model on the whole dataset. For instance, we can find subgroups where two target attributes have an unusual correlation, a classifier has a deviating predictive performance, or a Bayesian network fitted on several target attributes has an exceptional structure. We give an algorithmic

Responsible editor: M.J. Zaki.

This paper extends the previously published papers (Leman et al. 2008; Duivesteijn et al. 2010, 2012a).

W. Duivesteijn (✉)
Fakultät für Informatik, LS VIII, Technische Universität Dortmund, Dortmund, Germany
e-mail: wouter.duivesteijn@tu-dortmund.de

Ad J. Feelders
ICS, Utrecht University, Utrecht, the Netherlands
e-mail: a.j.feelders@uu.nl

A. Knobbe
LIACS, Leiden University, Leiden, the Netherlands
e-mail: knobbe@liacs.nl

solution for the EMM framework, and analyze its computational complexity. We also discuss some illustrative applications of EMM instances, including using the Bayesian network model to identify meteorological conditions under which food chains are displaced, and using a regression model to find the subset of households in the Chinese province of Hunan that do not follow the general economic law of demand.

Keywords Exceptional Model Mining · Subgroup Discovery · Supervised Local Pattern Mining · Regression · Bayesian Networks

Mathematics Subject Classification H.2.8: Data mining

1 Introduction

Traditionally, the goal of Subgroup Discovery (SD) (Klösgen 2002; Herrera et al. 2011) is to find interesting subsets of the dataset at hand. Commonly, a subset is deemed interesting when the distribution of one designated target attribute on the subset deviates from its distribution on the entire dataset. Consider for instance a dataset concerning people, and let the target attribute be whether the person develops lung cancer. Interesting subsets would then include the group of smokers, with an increased incidence of lung cancer, and the group of athletes, with a decreased incidence of lung cancer. SD explicitly requires a subset to have a concise description in terms of constraints on non-target attributes. A subset for which such a description exists is called a subgroup. This requirement makes subgroups easier to interpret by the data miner, but also more actionable for the domain expert who is interested in the real-world implications of the subgroup.

The notion of a deviating distribution of one designated target attribute does not encompass all forms of "interesting". We illustrate this with the synthetic example from Fig. 1. To our human eyes, it is immediately clear that the entire dataset (Fig. 1a), consists of a mixture of two phenomena: there is uniformly randomly distributed static (Fig. 1b), and there is something stronger going on along the diagonal (Fig. 1c). This diagonal forms a subset of the dataset where something interesting is going on, so ideally we would want our computer to find it for us in the dataset. It is not that easy for a computer to come up with this idea, for a variety of reasons:

- for each single record in the dataset, it is unclear whether it belongs to the diagonal or the static. Of course, when a record is in the top-left or bottom-right corner of Fig. 1a, it will surely be part of the static, but as we approach the diagonal it becomes less clear. After all, the dataset should be partitioned into a diagonal and uniformly randomly distributed static; it should not be partitioned into a diagonal and uniformly randomly distributed static *minus the diagonal*;
- for the partition to be *actionable*, we need a description for the diagonal in terms of other attributes in the dataset, in order to easily classify new records into one of the subsets and interpret the formed model in terms of the dataset domain;
- the model class is unknown. Our human eyes immediately see that there is a diagonal in the data, but I have to tell my computer what to look for. For all it knows, there is a parabola in the data, or a sinusoidal wave;

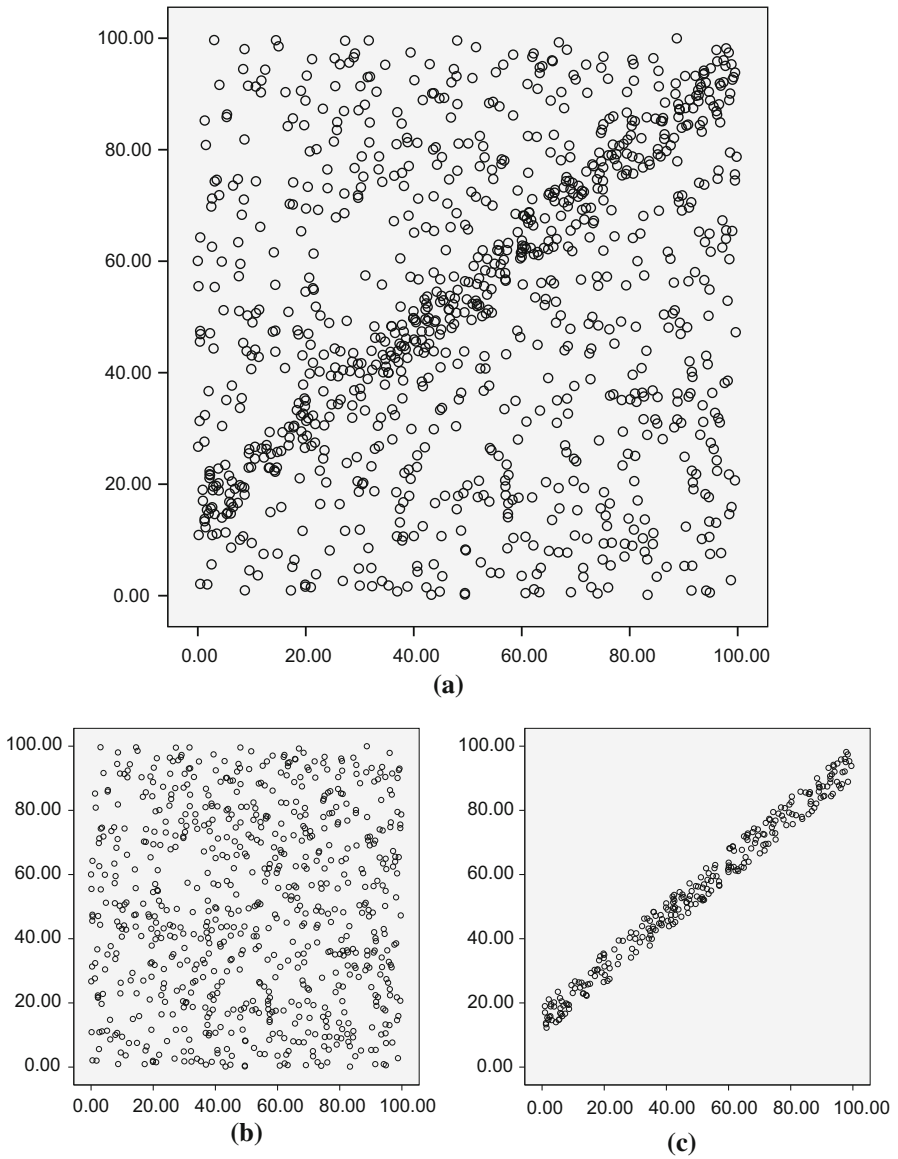

Fig. 1 A mixture of distributions. Ideally we would want to find a way to partition the original dataset from **a** into two parts: the static, depicted in **b**, and the diagonal line, depicted in **c**

– the model parameters are unknown. Even if we know that we want our computer to look for a diagonal, we still need to determine when we find such a diagonal really interesting. This is a matter of degree: do we find the diagonal in Fig. 1a pronounced enough to be reported? What if it becomes twice as thick? Obviously, if the diagonal would cover the entire dataset, we would not find it interesting; where do we draw the line?

In this paper, we do not solve all these issues. We do, however, develop a framework allowing users to define the model classes they are interested in, and to search for interesting subgroups. To capture this more general notion of interestingness in subgroups, we develop Exceptional Model Mining (EMM). It can be seen as a generalization of traditional SD that allows for more complicated target concepts. An EMM instance always starts with partitioning the attributes of the dataset into two sets: the descriptive attributes, which are used to define subgroups, and the target attributes, on which the subgroups are evaluated. A model class over these targets is selected, and a quality measure is defined that determines the interestingness of a subgroup in terms of model characteristics. Then, an SD run is performed to find subgroups with a high quality, i.e. having a model that is exceptional compared to the model on the entire dataset. Subgroup evaluation being based on any kind of characteristic of any kind of model, EMM discovers interesting subgroups for any interpretation of "interesting".

As this paper extends the previously published papers (Leman et al. 2008; Duivesteijn et al. 2010, 2012a) [the fourth publication Knobbe et al. (2012) is essentially a reproduction of Leman et al. (2008)], we detail in this paragraph which parts stem from which papers, and which parts are new. The Abstract, Introduction and Conclusion are all newly drafted. The motivating *Pisaster* example from Sect. 2 is taken from Duivesteijn et al. (2010), and the motivating *Giffen* example from Sect. 2 is taken from Duivesteijn et al. (2012a). The EMM framework definition in Sect. 3 is adapted from Leman et al. (2008), but the separation between descriptors and targets is made more explicit [as it is in Duivesteijn et al. (2010) and Duivesteijn et al. (2012a)]. The separation was not as strict in the EMM definition in Leman et al. (2008), and one can debate whether it is absolutely necessary. We think, however, that it is a healthy choice to make a clear separation between the attributes on which subgroups are formed and the attributes on which they are evaluated. The formal problem statement, discussion of our choices for the refinement operator and description language, and pseudocode for the beam search algorithm with accompanying complexity analysis (Sects. 3.1–4.2) are new contributions of this paper. Also new is the discussion in Sect. 3.2 on how to define an EMM instance, on common concepts in quality measures, and on whether a subgroup should be compared to its complement or to the whole dataset. The EMM instances from Sects. 5.1, 5.3, and 5.4 stem from Leman et al. (2008), the instance from Sect. 5.5 was introduced in Duivesteijn et al. (2010), and the instance from Sect. 5.6 stems from Duivesteijn et al. (2012a), while the Association model class from Sect. 5.2 is new. A similar division holds for the corresponding parts of Sect. 6: Sects. 6.1, 6.3, 6.4, 6.7.1, and 6.7.2 appeared in Leman et al. (2008), Sects. 6.5 and 6.7.3 stem from Duivesteijn et al. (2010), Sect. 6.6 is taken from Duivesteijn et al. (2012a), and Sect. 6.2 is new. Section 7.1 is relatively new: it obviously includes the related work discussed in Leman et al. (2008), Duivesteijn et al. (2010), Duivesteijn et al. (2012a), but it adds many more references, and categorizes and discusses them in more detail. The discussion of the three reasons why EMM exists in Sect. 8 is entirely new, including the subgroup-reinforced general linear regression modeling experiments of Sect. 8.1.

2 Motivation

Finding elements that behave differently from the norm in a dataset is a well-known task. Most data mining research in this direction focuses on *detecting* outliers: simply identifying the peculiarly-behaving records. The characteristic feature of local pattern mining techniques that separate them from outlier detection methods, is that in local pattern mining, we are not just looking for any outlying record or set of records in the data. Instead, we are looking for subgroups: coherent subsets that can be concisely described in terms of attributes of the data. The existence of such descriptions makes the subgroups more actionable. If we can tell a drug manufacturer that ten of his patients react badly to a certain type of medication, this doesn't help him much. However, if we can tell him that the group of smokers reacts badly, this gives the manufacturer a clear indication where to find a solution to his problem.

When the target concept in a dataset can no longer be captured by one particular attribute, but we still want to find exceptional subgroups in the dataset, we find a need for EMM. As an example of a relatively complex target concept, consider the research performed by Robert T. Paine in 1963 and 1964 in Makah Bay, WA (Paine 1966). 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 last three species.

In the studied ecosystem, *Pisaster* was the top carnivore: it consumed other species, but no other species consumed him, and *Pisaster* was the only species in the system for which both these statements held. This made Paine et al.'s research very relevant from a biological point of view; up until that point, it was generally assumed that removing the top carnivore from an ecosystem would increase diversity, but the *Pisaster* experiment proved that that was not necessarily the case.

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 directly 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*.

Paine models the food chains in the ecosystem as a Bayesian network. In order to find subgroups where the food chains between species are substantially different from the norm, we need to be able to detect the indirect processes that can be captured with a Bayesian network. Using an EMM instance, we can for instance find subgroups defined by environmental parameters in which complex food chains are displaced. The ability to cope with Bayesian networks makes the same EMM instance applicable

to datasets from such diverse fields as information retrieval (de Campos et al. 2004), gene expression in computational biology (Friedman et al. 2000), traffic accident reconstruction (Davis 2003), medical expert systems (Díez et al. 1997), and financial operational risk (Neil et al. 2005).

Another EMM instance could for example be used to find evidence for the Giffen effect in data. The economic law of demand states that, all else equal, if the price of a good increases, the demand for the product will decrease. Sir Robert Giffen described conditions under which this law does not hold (Marshall 1895). The classic example concerns extremely poor households, who mainly consume cheap staple food, and relatively rich households in the same neighborhood, who can afford to enrich their meals with a luxury food. In this situation, when the price of the staple food increases, there will be a point where the relatively rich households can no longer afford the luxury food. However, these people need to uphold their calorie intake. Hence, they react by consuming more of the cheapest food available to them, which is the staple food whose price just increased. For the relatively rich households in this poor neighborhood, an increase in the price of the staple food, will lead to an increase in the demand for the staple food. Notice that this relation does not hold for the extremely poor households: they consume only the staple food to begin with, so when the price increases they can simply afford less of it.

For a long time, the Giffen effect was a controversial theory in Economics, since no real-life dataset featuring the effect was available. However, in 2008, Jensen and Miller (2008) published a paper with the first real-world dataset containing the Giffen effect, for rice in Hunan, China. The relation between the price of and demand for certain goods is captured by a regression model. The group of relatively rich households in a poor neighborhood can be seen as a subgroup. Hence, an EMM instance mining for an unusual slope of a regression line can automatically detect such groups displaying Giffen behavior in a dataset.

3 The Exceptional Model Mining framework

We assume a dataset Ω to be a bag of N *records* $r \in \Omega$ of the form

$$r = \{a_1, \ldots, a_k, \ell_1, \ldots, \ell_m\}$$

where k and m are positive integers. We call $a_1, \ldots, a_k$ the *descriptive attributes* or *descriptors* of r, and $\ell_1, \ldots, \ell_m$ the *target attributes* or *targets* of r. The descriptors are taken from an unrestricted domain $\mathcal{A}$; restrictions on the type of each target may be imposed by the choice of model class. We refer to the ith record by r^i.

For our definition of subgroups, we need to define *descriptions*. These are functions $D : \mathcal{A} \to \{0, 1\}$. A description D *covers* a record r^i if and only if $D\left(a_1^i, \ldots, a_k^i\right) = 1$. Typically we restrict the *description language* $\mathcal{D}$ from which descriptions can be taken; the choice of description language within EMM is free. In Sect. 4.1 we detail the choice for $\mathcal{D}$ we make in this paper, but the Exceptional Model Miner is free to make another choice.

Definition 1 (*Subgroup*) A *subgroup* corresponding to a description D is the bag of records $G_D \subseteq \Omega$ that D covers:

$$G_D = \left\{ r^i \in \Omega \middle| D\left(a_1^i, \ldots, a_k^i\right) = 1 \right\}$$

From now on we omit the D if no confusion can arise, and refer to a subgroup as G. Whenever it is clear that we have a particular subgroup G in mind, we write n for the number of records in that subgroup: $n = |G|$. The complement of a subgroup is denoted by G^C, and for its number of records we write n^C. Hence, $G^C = \Omega \backslash G$, and $n^C = N - n$.

To objectively evaluate a candidate description in a dataset, we define a *quality measure*. For each description D in the description language $\mathcal{D}$, this function quantifies how exceptional the model is that we induce on G_D.

Definition 2 (*Quality Measure*) A *quality measure* is a function $\varphi : \mathcal{D} \rightarrow \mathbb{R}$ that assigns a unique numeric value to a description D, given a dataset Ω.

EMM (Leman et al. 2008; Duivesteijn et al. 2010, 2012a) is a data mining framework that can be seen as a generalization of the SD framework. SD strives to find descriptions that satisfy certain user-specified constraints. Usually these constraints include lower bounds on the quality of the description ($\varphi(D) \geq lb_1$) and size of the induced subgroup ($|G_D| \geq lb_2$). 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 description. Most common SD algorithms traverse[1] the search space of candidate descriptions in a general-to-specific way: they treat the space as a lattice whose structure is defined by a *refinement operator* $\eta : \mathcal{D} \rightarrow 2^{\mathcal{D}}$. This operator determines how descriptions can be extended into more complex descriptions by atomic additions. Most applications (including ours) assume η to be a *specialization operator*: $\forall\, D_i \in \eta(D_j) \;:\; D_j \succeq D_i$ (i.e. D_i is more specialized than D_j). The algorithm results in a ranked list of descriptions (or the corresponding subgroups) that satisfy the user-defined constraints.

In traditional SD, there is only a single target variable. Hence, the typical quality measure contains a component gauging the distributional difference of the target variable in the subgroup, 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 description is deemed interesting depends on both its exceptionality and the size of the corresponding subgroup.

EMM can be seen as an extension of SD. Rather than one single target variable, EMM uses a more complex target concept. An instance of EMM is defined by the combination of a chosen model class over the targets, and a designed quality measure over this model class. Having generated candidate subgroups to evaluate, for each subgroup under consideration we induce a model on the targets, learning the model from only the data belonging to the subgroup. Then, this model is evaluated with the

[1] We consider the exact search strategy to be a parameter of the algorithm.

designed quality measure, to determine which subgroups are the most interesting ones. The typical quality measure in EMM indicates how exceptional the model fitted on the targets in the subgroup is, compared to either the model fitted on the targets in its complement, or the model fitted on the targets in the whole dataset (we discuss this fundamental choice in Sect. 3.2.2). Just like in traditional SD, exceptional models are sometimes easily achieved in small subgroups, so if necessary an EMM quality measure also contains a component indicating the size of the subgroup.

3.1 Problem statement

So far, we have talked about EMM in an informal, colloquial manner. This is deliberate. The goal is to find interesting subgroups of a dataset, for whatever instantiation of "interesting" the user of EMM cares for, which is intrinsically subjective. Therefore, any formal definition of the EMM task will only concern a subset of what we attempt to achieve with EMM. Nevertheless, to provide a more precise handle on what we will be concerned with in the remainder of this paper, we can consider the following task definition.

Problem statement 1 (Top-q Exceptional Model Mining) Given a dataset Ω, description language $\mathcal{D}$, quality measure φ, positive integer q, and set of constraints $\mathcal{C}$, the Top-q EMM task is to find the list $\{D_1, \ldots, D_q\}$ of descriptions in the language $\mathcal{D}$ such that

- $\forall_{1 \leq i \leq q} : D_i$ satisfies all constraints in $\mathcal{C}$;
- $\forall_{i,j} : i < j \Rightarrow \varphi(D_i) \geq \varphi(D_j)$;
- $\forall_{D \in \mathcal{D} \setminus \{D_1, \ldots, D_q\}} : D$ satisfies all constraints in $\mathcal{C} \Rightarrow \varphi(D) \leq \varphi(D_q)$.

Informally, we find the q best-scoring descriptions in the description language that satisfy all constraints in $\mathcal{C}$. This set encompasses the limits to which we explore the search space, and potentially any other constraint that the Exceptional Model Miner would want to impose. In the Sect. 4, we discuss the choices made for the search space traversal and the refinement operator in the remainder of this paper. Note that the general EMM *framework* leaves the choice for these matters open.

Also noteworthy is the fact that this problem statement includes the traditional SD problem. This is a feature rather than a bug: we consider SD to be encompassed by EMM. In our view, SD is simply a version of EMM in which m, the number of targets, is set to 1.

3.2 How to define an EMM instance?

As previously described, an EMM instance is defined by the choice of model class over the targets, and quality measure over the model class. In Sect. 5, we define several such instances. Before that, we discuss some general themes that recur in EMM instance definitions.

The choice of model class is usually inspired by a real-life problem. For instance, to find conditions under which the expression of two genes interact in an unusual way,

one could choose a correlation model. When the goal is to find deviating dependencies between several species in an ecosystem, one is drawn towards graphical models. If we can formulate the relation between the targets for which we are interested in finding exceptions, this usually naturally directs our attention to a particular model class.

3.2.1 Quality measure concepts

Having chosen a model class, we need to define a quality measure that extracts characteristics from the learned models, and extracts from these characteristics a quantification of how different the models are from each other. Usually such a quantification is relatively straightforward to design. For instance, if the model class is a regression model with two variables, one could take the difference between the estimated slopes in each model as quality measure. However, such a quantification is typically not enough to design a proper measure for the quality of a description. After all, deviations from the norm are easily achieved in very small subsets of the data. Hence, directly taking a difference quantification as quality measure probably leads to descriptions of very small subgroups, which are usually not the most interesting ones to domain experts. Therefore, we somehow want to take the size of a subgroup into account in a quality measure.

In some of the canonical quality measures for SD, such as Weighted Relative Accuracy (WRAcc) (Lavrač et al. 1999), the size of a subgroup is directly represented by a factor n or $\sqrt{n}$. Though their simplicity is appealing, under certain circumstances one might argue that it is somewhat counter-intuitive to have a factor in a quality measure that explicitly favors subgroups covering the entire dataset over smaller subgroups. A slightly more sophisticated way to represent the subgroup size, is to multiply (i.e. weigh) the quantification of model difference with the *entropy* of the split between the subgroup and its complement. The entropy captures the information content of such a split, and favors balanced splits (1 bit of information for a 50/50 split) over skewed splits (0 bits for the extreme case of either subgroup or complement being empty). The entropy function $\varphi_{\text{ef}}(D)$ is defined (in this context) as

$$\varphi_{\text{ef}}(D) = -n/N \lg n/N - n^C/N \lg n^C/N$$

Another way to direct the search away from extremely small subgroups, is by employing a quality measure based on a statistical test. For certain models there may be hypotheses of the form

H_0 : model parameter for description = model parameter for complement;
H_1 : model parameter for description $\neq$ model parameter for complement.

If statistical theory enables us to compute a p value corresponding to this test, then we could use $1 - p$ as the quality measure. Hence, we have constructed a measure ranging from 0 to 1 for which higher values indicate more interesting descriptions. *Notice that we do not employ these p values to assess the statistical significance of the found subgroups.* In a typical mining run, a vast number of candidate subgroups is

considered. Hence, to properly assess significance, one should deal with the multiple comparisons problem (Hochberg and Tamhane 1987): the p values should at least be corrected to assess whether a subgroup could be considered statistically significant. This, however, is not our goal. We merely employ the p values as components in a statistically inspired quality measure: a function that is meant to compare the relative merits of subgroups, but not to reject some and accept others. Alternatively but equivalently, we could use the test statistic upon which the p value computation is based as the quality measure.

Sections 5.1 (φ_{scd}), 5.4 (φ_{sed}) and 5.3 (φ_{ssd}) feature examples of quality measures that are directly based on a statistical test. In Sects. 5.1.1 (φ_{ent}) and 5.5 (φ_{weed}) we find examples of quality measures employing the entropy function. Quality measures from Sects. 5.1.1 (φ_{abs}), 5.4.1 (φ_{BDeu} and φ_{Hel}), 5.5.1 (φ_{ed}), and 5.6 (φ_{Cook}) consist solely of a difference quantification (occasionally these are statistically inspired, but they are not directly based on an established statistical test).

Notice that the choice of quality measure is left entirely to the whims and fancies of the Exceptional Model Miner: we find it important to let our framework allow the user to specify exactly what he/she finds exceptional. That being said, in our opinion it would generally make sense when quality measures incorporate the abovementioned concepts. As a rule of thumb, we would advise to strive for statistically based quality measures. However, there may very well be reasons to forego these, for instance when they are computationally too expensive to be incorporated into EMM, or when they depend on assumptions that we do not want to impose on our subgroups. In those cases, our rule of thumb would be to look beyond the statistically inspired measures, and consider the other categories. In Sect. 6.7 we discuss some subgroups found with alternative quality measures for particular model classes, allowing a comparison with the main quality measure we propose for the same model class.

3.2.2 Compared to what?

So far we have discussed quality measure development as a means of assessing how different two learned models are from one another, and how to ensure that subgroups have a substantial size. However, we have neglected a cardinal point. Since a quality measure should assign a quality to a description, its model should be compared, but to which other model? There are two options: we can compare the model for a description of a subgroup G either to the model for its complement G^C, or to the model for the whole dataset Ω. The simple constructed example from Fig. 2 illustrates that these two comparisons can lead to very different outcomes.

Suppose that we have a two-dimensional target space, and we are concerned with finding descriptions having a deviating regression line in these two dimensions. Figure 2 depicts the target space, and the six records in the example dataset. The dotted grey line is the regression line of the whole dataset, with slope -1. Now suppose that we find the description D covering the records depicted as circles. The dashed grey line is the regression line of G_D, with slope 1. The solid grey line is the regression line of G_D^C, also having slope 1. When gauging the exceptionality of a description solely by the slope of the regression line, we find G_D interesting when compared to Ω, but

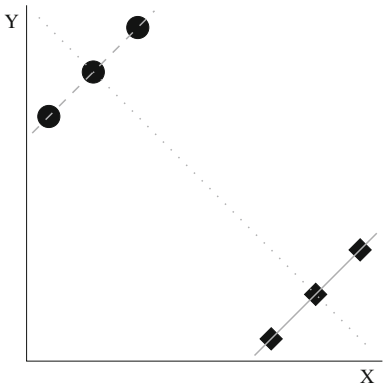

Fig. 2 Should we compare a subgroup G to its complement G^C, or to the whole dataset Ω?

not at all when compared to G_D^C. Of course, the assessment changes when we include the intercept in the evaluation.

The problem as displayed in Fig. 2 is underdetermined; we do not have enough information to formulate an opinion on whether the subgroup should be deemed interesting. It can therefore not be used to illustrate whether comparing to G_D^C or to Ω is preferable; it merely illustrates that a different choice may lead to a different outcome.

There is not always a clear-cut preferred choice whether to compare to G_D^C or to Ω. Sometimes, the real-life problem at hand can point in one direction: if we are interested in deviations from a possibly inhomogeneous norm, it makes more sense to compare to Ω, whereas if we are interested in dichotomies, it makes more sense to compare to G_D^C. On other occasions, a statistically inspired quality measure may *require* choosing either Ω or G_D^C, to prevent violation of mathematical assumptions. Lastly, when the model class is so complicated that learning models from data covered by descriptions has a nontrivial computational expense, efficiency might dictate the choice: when comparing n descriptions to Ω, learning $n+1$ models suffices, but when comparing them to G_D^C, learning $2n$ models is required. In our view, there is no general correct choice what to compare to. It is important for an Exceptional Model Miner to realize that, at least in theory, this choice can fundamentally influence the outcome.

4 Algorithmic solution

Since the goal of SD/EMM is to find interesting subsets of the data, the corresponding search space could potentially be exponentially large in the number of records.[2] Hence, we cannot simply explore this space by brute force; we need to find a more sophisticated search strategy. Usually, part of the problem is already solved by only allowing subgroups. Since subgroups are subsets of the data for which a description exists, the set of subgroups is typically smaller than the set of subsets. Unfortunately, when many descriptors in the dataset are numeric, the difference is not very large.

[2] When the description language at hand is very expressive, and the dataset contains many numeric attributes, one can imagine that for every subset of the dataset at least one corresponding description exists.

There are two main schools of thought in the community on how to overcome this problem, each with their own focus. The one, following classical SD papers (Wrobel 1997; Lavrač et al. 1999), restricts the attributes in the dataset to be nominal and imposes an anti-monotonicity constraint on the used quality measure. Then the resulting search space can be explored exhaustively. The other resorts to heuristic search. This allows the attributes to be numeric as well, and facilitates a general quality measure. Since EMM is developed to capture any concept of interestingness in subgroups, we value the capacity for handling any quality measure and numeric attributes over exhaustiveness. Hence we select the heuristic path. Exhaustive SD methods and alternative search strategies are discussed in further detail in Sect. 7.1.

In the EMM setting, usually the *beam search* strategy is chosen, which performs a level-wise search. On each level, the best w descriptions according to our quality measure φ are selected, and refined to create the candidate descriptions for the next level. The search is constrained by an upper bound on the complexity of the description 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 descriptions at each level keeps the process focused and thus more tractable.

4.1 Refinement operator and description language

An important part of the beam search strategy is generating the set of candidate descriptions for the next level, by refining descriptions on the current level. This process is guided by the refinement operator η and the description language $\mathcal{D}$, for which we detail our choices in this section. Our description language $\mathcal{D}$ consists of logical conjunctions of conditions on single attributes.

We treat the numeric attributes with a particular kind of discretization, starting by fixing a positive integer $b \leq N$ (the number of *bins*) before the EMM process starts. On the first search level, when the generating description has no conditions, the discretization we apply is equal to static pre-algorithm discretization of the attribute into b bins of equal frequency. However, on each subsequent search level, our generating descriptions consist of a positive number of conditions, hence they cover strictly less than N records. Since on these levels we consider a discretization into b equal-frequency bins of the attribute-values *within the generating non-empty description*, the bins may be different for each generating description. This *dynamic discretization* during the process draws more information from the attribute than we would get when statically discretizing it beforehand. Notice that this type of discretization is a general technique, rather than being endemic to EMM.

When η is presented with a description D to refine, it builds up the set $\eta(D)$ by looping over all the descriptive attributes $a_1, \ldots, a_k$. For each attribute, a number of descriptions are added to $\eta\,(D)$, depending on the attribute type

if a_i is binary: add $D \cap (a_i = 0)$ and $D \cap (a_i = 1)$ to $\eta(D)$;
if a_i is nominal, with values $v_1, \ldots, v_g$: add $\bigcup_{j=1}^{g} \{D \cap (a_i = v_j),\, D \cap (a_i \neq v_j)\}$ to $\eta(D)$;

if a_i is numeric: order the values of a_i that are covered by the description D; this gives us a list of ordered values $v_{(1)}, \ldots, v_{(n)}$ (where $n = |G_D|$). From this list we select the split points $s_1, \ldots, s_{b-1}$ by letting

$$s_j = v_{(\lfloor j\frac{n}{b} \rfloor)}$$

Then, add $\left\{ D \cap (a_i \leq s_j),\ D \cap (a_i \geq s_j) \right\}_{j=1}^{b-1}$ to $\eta(D)$.

Informally, when presented with a description D, η builds a set of refinements by considering the descriptive attributes one by one. Each such refinement consists of the conditions already present in D, plus one new condition. If an encountered attribute a_i is binary, 2 refined descriptions are added to $\eta(D)$: one for which D holds and a_i is true, and one for which D holds and a_i is false. If the attribute a_i is nominal with g different values, $2g$ refined descriptions are added to $\eta(D)$: for each of the g values, one where D holds and the value is present, and one where D holds and any of the $g-1$ other values is present. If the attribute a_i is numeric, we divide the values for a_i *that are covered by* D into a predefined number b of equal-sized bins. Then, using the $b-1$ *split points* $s_1, \ldots, s_{b-1}$ that separate the bins, $2(b-1)$ refined descriptions are added to $\eta(D)$: for each split point s_j, one where D holds and a_i is less than or equal to s_j, and one where D holds and a_i is greater than or equal to s_j.

4.2 Beam search algorithm for Top-q Exceptional Model Mining

Having described our choices for the search strategy and refinement operator that we use in the remainder of this paper, we can now describe and analyze an algorithm for the top-q EMM problem stated in Sect. 3.1. The pseudocode is given in Algorithm 1. In the algorithm, we assume that there is a subroutine called SATISFIESALL that tests whether a candidate description satisfies all constraints in a given set. Among the abstract datastructures we assume, the Queue is a standard queue with unbounded length. The PriorityQueue(x) is a queue containing at most x elements, where elements are stored and sorted with an associated quality; only the x elements with the highest qualities are retained, while other elements are discarded. In a straightforward but not too naive implementation, a PriorityQueue is built with a heap as its backbone. In this case the elementary operations, *insert_with_priority* for adding an element to the PriorityQueue and *get_front_element* for removing the element with the highest quality from the PriorityQueue, have a computational cost of $\mathcal{O}(\log x)$ (Knuth 1998, pp. 148–151).

Many statements in the algorithm control the beam search process in a straightforward manner. However, the process is also controlled by the interplay between the different (Priority-)Queues, which is more intricate and deserves attention. The resultSet is a PriorityQueue maintaining the q best descriptions found so far. Nothing is ever explicitly removed from the resultSet, but if the quality of a description is no longer among the q best, it is automatically discarded. Hence, the resultSet maintains the final result that we seek. The beam is a similar PriorityQueue, but with a different role. Here, the w best descriptions found so far *on the current search level* are maintained. When all candidates for a search level have been explored, the contents of the

Algorithm 1 Beam Search for Top-q Exceptional Model Mining

Input: Dataset Ω, quality measure φ, refinement operator η, beam width w, beam depth d, result set size q, Constraints $\mathcal{C}$
Output: PriorityQueue resultSet

```
1 : candidateQueue ← new Queue;
2 : candidateQueue.enqueue({});                          ▷ Start with empty description
3 : resultSet ← new PriorityQueue(q);
4 : for (Integer level ← 1; level ≤ d; level++) do
5 :     beam ← new PriorityQueue(w);
6 :     while (candidateQueue ≠ ∅) do
7 :         seed ← candidateQueue.dequeue();
8 :         set ← η(seed);
9 :         for all (desc ∈ set) do
10 :            quality ← φ(desc);
11 :            if (desc.SATISFIESALL(C)) then
12 :                resultSet.insert_with_priority(desc,quality);
13 :                beam.insert_with_priority(desc,quality);
14 :    while (beam ≠ ∅) do
15 :        candidateQueue.enqueue(beam.get_front_element());
16 : return resultSet;
```

beam are moved into the unbounded but (by then) empty Queue candidateQueue, to generate the candidates for the next level.

4.2.1 Complexity

Since EMM is a highly parametrized framework, instantiated by a model class and quality measure, we need to introduce some notation before we can analyze the computational complexity of the algorithm. We write $M(n, m)$ for the cost of learning a model from n records on m targets, and c for the cost of comparing two models from the chosen model class.

Theorem 1 *The worst-case computational complexity of Algorithm 1 is*

$$\mathcal{O}\left(dwkN\left(c + M(N, m) + \log(wq)\right)\right)$$

Proof We start our analysis at the innermost loop, working bottom-up. Line 12 inserts an element into a PriorityQueue of size q (the number of subgroups the algorithm should report), which costs $\mathcal{O}(\log q)$. Line 13 does the same for a PriorityQueue of size w (the beam width), and hence costs $\mathcal{O}(\log w)$. The conditions checked in line 11 are the user-induced constraints a domain expert may impose on the resulting descriptions. These usually are relatively simple conditions concerning for instance a minimal number of records covered by the descriptions. As such, they are relatively cheap to check. For all reasonable constraints a domain expert may come up with, the necessary information can be extracted during the same scans of the dataset we need when, for instance, computing the quality of the description in the preceding line. As such, we assume the computational complexity of line 11 to be dominated by the complexity of line 10. The worst-case scenario is that all descriptions pass the test, hence the commands inside the if-clause need to be computed every time. Thus, the total complexity of lines 11 through 13 is $\mathcal{O}(\log w + \log q) = \mathcal{O}(\log(wq))$.

Line 10 computes the quality of a description. In the worst case, this requires the learning of two models: one on the description and one on its complement, and comparing these models. Hence: $\mathcal{O}(c + 2M(N, m)) = \mathcal{O}(c + M(N, m))$ (recall the definition of c and $M(N, m)$, as introduced just before Theorem 1). In line 9, a loop is run for all refinements of a seed description. By our choice of refinement operator η, the worst case would be if every descriptive attribute were nominal (or numeric) having N (the number of records in the dataset) distinct values. For each of the k descriptors, we would then generate $2N$ refinements. The loop is thus repeated $2kN$ times, which costs $\mathcal{O}(kN)$. Hence, the total complexity of lines 9 through 13 is $\mathcal{O}(kN(c + M(N, m) + \log(wq)))$.

Line 8 enumerates all refinements of one description, which we have just analyzed to cost $\mathcal{O}(kN)$. Line 7 dequeues an element from an ordinary Queue, which can be done in $\mathcal{O}(1)$. Line 6 loops all previously analyzed lines as many times as there are elements in the candidateQueue. This queue never has more than w elements, since it is always emptied before (in line 15) at most w new elements are added to the queue. Hence, the total complexity of lines 6 through 13 is $\mathcal{O}(w(kN + kN(c + M(N, m) + log(wq)))) = \mathcal{O}(wkN(c + M(N, m) + \log(wq)))$.

On the same level we find line 5, which costs $\mathcal{O}(1)$, and the while-loop of lines 14 through 15, which costs $\mathcal{O}(w \log w)$ if done extremely naively. These lines are dominated in complexity by lines 6 through 13. All these lines are enveloped by a for-loop starting at line 4, which is repeated d (the beam depth) times. Lines 1 through 3 and 16 can be computed in constant time, and so the total computational complexity of Algorithm 1 becomes

$$\mathcal{O}(dwkN(c + M(N, m) + \log(wq)))$$

□

This complexity seems relatively benign; we see no factors with exponents higher than one, and the worst parameter has complexity $\mathcal{O}(w \log w)$, which is tractable for a generous range of values for w. However, there are some variables in the complexity expression, which can lead to higher powers of parameters if we fill them in by selecting a model class and quality measure. For instance, if we would perform traditional SD with this algorithm, we would be searching for descriptions having an unusually high mean for one designated target. Hence, the model computation complexity becomes $M(N, 1) = \mathcal{O}(N)$, and the model comparison cost becomes $c = \mathcal{O}(1)$. Thus, the total computational complexity of Beam Search for Top-q SD would be $\mathcal{O}(dwkN(N + \log(wq)))$, which is quadratic in the number of records in the dataset.

Note that this computational complexity is in many respects a worst-case scenario, whose bounds a real-life run of the algorithm is unlikely to meet. Since data of such high cardinality is rarely obtained, the number of refinements of a seed description is usually much lower than $2kN$. Also, unlike in the worst-case scenario, the beam search converges in such a way that per search level the subgroups reduce in size, hence the modeling is done over progressively smaller parts of the dataset. Also noteworthy are the facts that when a dataset is extended with more data of the same cardinality, the algorithm scales linearly, and that the number of candidates under consideration is roughly equal per search level, except for level $d = 1$.

5 Exceptional Model Mining instances

In this section we define several sensible model classes, and develop appropriate quality measures for each of them. In several of these model classes, notational conventions hold to which we strive to adhere. Consequentially, we will overload some symbols with multiple meanings across the following subsections. These multiply-defined symbols reappear in the same context in the respective subsections of Sect. 6. Please be careful not to confuse the symbols over different model classes; they are kept consistent and unambiguous *within* every model class.

5.1 Correlation

In the correlation model, we consider two numeric targets, ℓ_1 and ℓ_2. Within the correlation model class, we refer to them as $x = \ell_1$ and $y = \ell_2$. We are interested in their linear association as measured by the correlation coefficient ρ. We estimate ρ by the sample correlation coefficient $\hat{r}$:

$$\hat{r} = \frac{\sum \left(x^i - \bar{x}\right)\left(y^i - \bar{y}\right)}{\sqrt{\sum \left(x^i - \bar{x}\right)^2 \sum \left(y^i - \bar{y}\right)^2}}$$

where x^i denotes the ith observation on x, and $\bar{x}$ denotes its mean. We let ρ^G and ρ^{G^C} denote the population coefficients of correlation for G and G^C, respectively, and let $\hat{r}^G$ and $\hat{r}^{G^C}$ denote their sample estimates.

To find descriptions with a substantial coverage and deviating correlation coefficient, we develop a statistically-oriented quality measure, based on the test

$$H_0 : \rho^G = \rho^{G^C} \quad \text{against} \quad H_1 : \rho^G \neq \rho^{G^C}$$

Generally, the sampling distribution of $\hat{r}$ is not known. If x and y follow a bivariate normal distribution, then we can apply the Fisher z transformation

$$z' = \frac{1}{2} \ln \left(\frac{1 + \hat{r}}{1 - \hat{r}} \right)$$

The sampling distribution of z' is approximately normal (Neter et al. 1966).

As a consequence,

$$z^* = \frac{z' - z^{C'}}{\sqrt{\frac{1}{n-3} + \frac{1}{n^C - 3}}}$$

approximately follows a standard normal distribution under H_0.

If both n and n^C are greater than 25, then the normal approximation is quite accurate, and can safely be used to compute the p values. As quality measure φ_{scd} we take 1 minus the computed p value. Because we have to introduce the normality assumption

to be able to compute the p values, φ_{scd} should be viewed as a heuristic measure. Transformation of the original data (for example, taking their logarithm) may make the normality assumption more reasonable.

5.1.1 Alternatives

A logical consideration for a quality measure would be the absolute difference of the correlation for the description D and its complement, i.e.

$$\varphi_{\text{abs}}(D) = \left| \hat{r}^{G_D} - \hat{r}^{G_D^C} \right|$$

Unfortunately, this measure does not take into account the coverage of the descriptions, and hence does not do anything to prevent overfitting.

As an improvement of φ_{abs}, the following quality function weighs the absolute difference between the correlations with the *entropy function* of the split between the description and its complement, as introduced in Sect. 3.2.1. Hence, when we find descriptions with φ_{abs}, but we find their coverage not sufficient, we can solve this problem by running EMM with the alternative quality measure φ_{ent}, defined as

$$\varphi_{\text{ent}}(D) = \varphi_{\text{ef}}(D) \cdot \left| \hat{r}^{G_D} - \hat{r}^{G_D^C} \right|$$

5.2 Association

In analogy to the correlation model between two numeric targets, we turn to the association between two nominal targets, denoted by $x = \ell_1$ and $y = \ell_2$. Let the values of x be coded as $0, 1, \ldots, d_x - 1$, where d_x is the cardinality of x. The values of y are coded in a similar fashion. Even though we use integers to code the distinct values of x and y, their values are treated as unordered.

We are interested in finding subgroups where the association between x and y is markedly different from their association in the complement. To this end, we propose to compare two log-linear models (Goodman 1970) on the set of variables x, y, and D, where D is the description inducing the subgroup we strive to evaluate. These models are the saturated model (allowing the association between x and y to be different for $D = 0$ and $D = 1$), and the so-called homogeneous association model (enforcing the constraint that the association between x and y is the same within the subgroup and its complement). In terms of their log-linear expansion, the two models are:

$$\begin{aligned}
\log P(x, y, D) = {} & u_{\varnothing} + u_x(x) + u_y(y) + u_D(D) && \text{(saturated model)} \\
& + u_{xy}(x, y) + u_{xD}(x, D) + u_{yD}(y, D) \\
& + u_{xyD}(x, y, D) \\
\log P(x, y, D) = {} & u_{\varnothing} + u_x(x) + u_y(y) + u_D(D) && \text{(homogeneous association)} \\
& + u_{xy}(x, y) + u_{xD}(x, D) + u_{yD}(y, D)
\end{aligned}$$

Here $u_{xy}(x, y)$ is called the u-term associated with the variable pair x, y. To avoid getting too many parameters in the model, $u_{xy}(x, y)$ is constrained to be zero when at

least one of x and y has the value zero. Analogous constraints are applied to the other u-terms. In this way, the saturated model has $d_x \times d_y \times 2$ u-terms, that is, as many as there are cells in the three-way contingency table of x, y and D. For further details we refer the reader to Chap. 7 of Whittaker (1990).

The difference between these two models is the absence of the three-way interaction term $u_{xyD}(x, y, D)$ in the homogeneous association model. Because this three-way interaction term is constrained to be zero, the homogeneous association model cannot model the case where the association between x and y differs between the subgroup and its complement.

The deviance of the homogeneous association model is commonly used as a test statistic for comparison against the saturated model. Therefore we use this quantity as a quality measure, where a high deviance corresponds to a high quality description. The deviance corresponding to description D is given by:

$$\varphi_{\text{dev}}(D) = 2 \sum_{\text{cells}} \text{observed}(D) \cdot \log \frac{\text{observed}(D)}{\text{fitted}(D)}$$

In this expression, the summation ranges over all cells of the contingency table of x, y and D, *observed* refers to the observed count of a cell, and *fitted* refers to the maximum likelihood fitted counts of the homogeneous association model. It should be noted that this quality measure does not have a natural protection against overfitting, so it should be used in combination with some minimal support threshold on the subgroup size. To use the deviance as a basis for testing, the chi-squared approximation to the distribution of the deviance is commonly regarded as accurate if the fitted counts are at least 5 for each cell of the contingency table (Agresti 1990). A similar minimum support restriction can be employed to use the deviance as a quality measure.

5.3 Simple linear regression

In this section, we discuss some possibilities of EMM with simple regression models, allowing only one output ($y = \ell_2$) and one input variable ($x = \ell_1$) in the regression:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{1}$$

We will discuss a more general linear regression model in Sect. 5.6.

Consider model (1) fitted to a subgroup G and its complement G^C. Of course, there is a choice of distance measures between the fitted models. We propose to look at the difference in the slope β_1 between the two models, because this parameter is usually of primary interest when fitting a regression model: it indicates the change in the expected value of y, when x increases with one unit. Another possibility would be to look at the intercept β_0, if it has a sensible interpretation in the application concerned. Like with the correlation coefficient, we use significance testing to measure the distance between the fitted models. Let β_1^G be the slope for the regression function of G and $\beta_1^{G^C}$ the slope for the regression function of G^C. The hypothesis to be tested is

$$H_0 : \beta_1^G = \beta_1^{G^C} \quad \text{against} \quad H_1 : \beta_1^G \neq \beta_1^{G^C}$$

We use the least squares estimate $\hat{\beta}_1$ for the slope β_1, and unbiased estimator s^2 for the variance of $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})\,(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \qquad s^2 = \frac{\sum \hat{\varepsilon}_i^2}{(\xi - 2)\sum (x_i - \bar{x})^2}$$

where $\hat{\varepsilon}_i$ is the regression residual for individual i, and ξ is the sample size. Finally, we define our test statistic t'. Although it does not have a t distribution, its distribution can be approximated quite well by one, with degrees of freedom given below (cf. Moore and McCabe (1993)):

$$t' = \frac{\hat{\beta}_1^G - \hat{\beta}_1^{G^C}}{\sqrt{s^{G^2} + s^{G^C 2}}} \qquad df = \frac{\left(s^{G^2} + s^{G^C 2}\right)^2}{\frac{s^{G^4}}{n-2} + \frac{s^{G^C 4}}{n^C - 2}}$$

The approximation is accurate when $n + n^C \geq 40$, so unless we analyze a very small dataset we should be confident to base p value computation on it. Our quality measure φ_{ssd} is one minus this p value.

5.4 Classification

In the case of classification, we are dealing with models for which the output attribute $y = \ell_m$ is discrete. In general, the attributes $\ell_1, \ldots, \ell_{m-1}$ can be of any type (binary, nominal, numeric, etc). Furthermore, our EMM framework allows for any classification method, as long as some quality measure can be defined in order to judge the models induced. Although we allow arbitrarily complex methods, such as decision trees, support vector machines or even ensembles of classifiers, we only consider a relatively simple classifier here, for reasons of simplicity and efficiency.

Analogous to the linear regression case, we consider the logistic regression model

$$\text{logit}(P(y_i = 1|x_i)) = \ln\left(\frac{P(y_i = 1|x_i)}{P(y_i = 0|x_i)}\right) = \beta_0 + \beta_1 \cdot x_i$$

where $y \in \{0, 1\}$ is a binary class label and $x \in \{\ell_1, \ldots, \ell_{m-1}\}$. The coefficient β_1 tells us something about the effect of x on the probability that y occurs, and hence may be of interest to subject area experts. A positive value for β_1 indicates that an increase in x leads to an increase of $P(y = 1|x)$. The strength of influence can be quantified in terms of the change in the odds of $y = 1$ when x increases with, say, one unit.

To judge whether the effect of x is substantially different in a particular subgroup G_D (built from a description D), we fit the model

$$\text{logit}(P(y_i = 1|x_i)) = \beta_0 + \beta_1 \cdot D(i) + \beta_2 \cdot x_i + \beta_3 \cdot (D(i) \cdot x_i) \tag{2}$$

where $D(i)$ is shorthand notation for $D\left(a_1^i, \ldots, a_k^i\right)$ Note that

$$\text{logit}(P(y_i = 1|x_i)) = \begin{cases} (\beta_0 + \beta_1) + (\beta_2 + \beta_3) \cdot x_i & \text{if } D(i) = 1 \\ \beta_0 + \beta_2 \cdot x_i & \text{if } D(i) = 0 \end{cases}$$

Hence, we allow both the slope and the intercept to be different in the subgroup and its complement. As a quality measure, we propose to use one minus the p value of a test on $\beta_3 = 0$ against a two-sided alternative in the model of Eq. (2). This is a standard test in the literature on logistic regression (Neter et al. 1966). We refer to this quality measure as φ_{sed}.

5.4.1 Alternatives

Another classifier we can consider is the *Decision Table Majority* (DTM) classifier (Kohavi 1995), also known as a *simple decision table*. The idea behind this classifier is to compute the relative frequencies of the ℓ_m values for each possible combination of values for $\ell_1, \ldots, \ell_{m-1}$. For combinations that do not appear in the dataset, the relative frequency estimates are based on that of the whole dataset. The predicted ℓ_m value for a new individual is simply the one with the highest probability estimate for the given combination of input values.

Example 1 As an example of a DTM classifier, consider a hypothetical dataset of 100 people applying for a mortgage. The dataset contains two attributes describing the age (divided into three suitable categories) and marital status of the applicant. A third attribute indicates whether the application was successful, and is used as the output. Out of the 100 applications, 61 were successful. The following decision table lists the estimated probabilities of success for each combination of *age* and *married*. The support for each combination is indicated between brackets.

	Married = 'no'	Married = 'yes'
Age = 'low'	0.25 (20)	0.61 (0)
Age = 'medium'	0.4 (15)	0.686 (35)
Age = 'high'	0.733 (15)	1.0 (15)

As this table shows, the combination *married* = 'yes' $\wedge$ *age* = 'low' does not appear in this particular dataset, and hence the probability estimate is based on the complete dataset (0.61). This classifier predicts a positive outcome in all cases except when *married* = 'no' and *age* is either 'low' or 'medium'.

For this instance of the classification model, we discuss two different quality measures. The *Bayesian Dirichlet equivalent uniform* (BDeu) score, which can be used as a measure for the performance of the DTM classifier on G_D, and the *Hellinger distance*, which assigns a value to the distance between the conditional probabilities estimated on G_D and G_D^C.

BDeu Score (φ_{BDeu}) The BDeu score φ_{BDeu} is a measure from Bayesian theory (Heckerman et al. 1995) and is used to estimate the performance of a classifier for a description, with a penalty for small contingencies that may lead to overfitting. Note that this measure ignores how the classifier performs on the complement. It merely captures how "predictable" a particular description is.

The BDeu score is defined as

$$\prod_{\ell_1,\ldots,\ell_{m-1}} \frac{\Gamma(\theta/\mathcal{I})}{\Gamma(\theta/\mathcal{I} + \#(\ell_1,\ldots,\ell_{m-1}))} \prod_{\ell_m} \frac{\Gamma(\theta/\mathcal{I}\mathcal{J} + \#(\ell_1,\ldots,\ell_m))}{\Gamma(\theta/\mathcal{I}\mathcal{J})}$$

where Γ denotes the gamma function, $\mathcal{I}$ denotes the number of value combinations of the input variables, $\mathcal{J}$ the number of values of the output variable, and $\#(\ell_1,\ldots,\ell_m)$ denotes the number of records with that value combination. The parameter θ denotes the *equivalent sample size*. Its value can be chosen by the user.

Hellinger (φ_{Hel}) Another possibility is to use the Hellinger distance (Yang and Le Cam 2000). It defines the distance between two probability distributions $P(x)$ and $Q(x)$ as follows

$$H(P, Q) = \sum_x \left(\sqrt{P(x)} - \sqrt{Q(x)}\right)^2$$

where the sum is taken over all possible values x. In our case, the distributions of interest are

$$P\left(\ell_m \mid \ell_1,\ldots,\ell_{m-1}\right)$$

for each possible value combination $\ell_1,\ldots,\ell_{m-1}$. The overall distance measure becomes

$$\begin{aligned}\varphi_{\text{Hel}}(D) &= H\left(\hat{P}^{G_D}, \hat{P}^{G_D^C}\right) \\ &= \sum_{\ell_1,\ldots,\ell_{m-1}} \sum_{\ell_m} \left(\sqrt{\hat{P}^{G_D}(\ell_m|\ell_1,\ldots,\ell_{m-1})} - \sqrt{\hat{P}^{G_D^C}(\ell_m|\ell_1,\ldots,\ell_{m-1})}\right)^2\end{aligned}$$

where $\hat{P}^{G_D}$ denotes the probability estimates on G_D. Intuitively, we measure the distance between the conditional distribution of ℓ_m in G_D and G_D^C for each possible combination of input values, and add these distances to obtain an overall distance.

5.5 Bayesian network

The Bayesian network model was inspired by the *Pisaster* example from Sect. 2. We consider multiple nominal targets $\ell_1,\ldots,\ell_m$. A subgroup is deemed interesting in this setting, when the conditional dependence relations between the targets are substantially different on the subgroup than these relations on the whole dataset. Hence we validate

the subgroups on the interdependencies between the targets, rather than the target values themselves. To capture these interdependencies, we learn a Bayesian network between the targets, from data.

There are many ways in which a Bayesian network can be learned from data. We use a non-deterministic hill climbing algorithm, but in no way is this essential to the method presented here. The choice of method can be considered a parameter of this EMM instance, hence we consider the details of the selected method to be irrelevant to the reader of this paper. Feel free to plug in your own preferred method. We refer the reader who is interested in the details to Sect. III.B of Duivesteijn et al. (2010).

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 learn a Bayesian network BN_Ω between the targets from the entire 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 subgroup G_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. Instead, we propose a heuristic quality measure based on the following well-known result by Verma and Pearl (1990):

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 differences in skeletons and v-structures as a measure of how different two DAGs are.

Definition 3 (*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

$$\zeta = \Big|[S_1 \ominus S_2] \ \cup \ [M_1 \ominus M_2]\Big|$$

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 an exclusive disjunction: $X \ominus Y = (X \cup Y) - (X \cap Y)$. 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.

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

Definition 4 (*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:

$$\varphi_{\text{ed}}(D) = \delta\left(BN_\Omega, BN_D\right)$$

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} from Sect. 3.2, to obtain the following aggregate measure.

Definition 5 (*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 to reduce its bias towards 50/50 splits, since we are primarily interested in a subgroup with large edit distance, while mediocre entropy is acceptable.

5.5.1 Alternatives

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. Alternatively, one could divert one's attention away from the learned structure of the model, and focus on the underlying joint probability distribution of the Bayesian network models. Having computed the joint probability distribution for both a subgroup and either its complement or the whole dataset, one could then for instance employ the Hellinger distance φ_{Hel} from Sect. 5.4.1.

5.6 General linear regression

In this section, we investigate the more general case of the model investigated in Sect. 5.3: a linear regression model on multiple target attributes $\ell_1, \ldots, \ell_m$. We learn the linear regression model

$$Y = X\beta + \varepsilon$$

where Y is the $N \times 1$ vector of ℓ_m-values from our dataset, X is the $N \times m$ full rank matrix of which column 1 consists of N times the value 1 and column $i+1$ consists of the ℓ_i-values from our dataset (with $i \in \{1, \ldots, m-1\}$), β is the unknown $m \times 1$ vector consisting of the regression parameters, and ε is the $N \times 1$ vector of randomly distributed errors such that $\mathbb{E}(\varepsilon) = 0$ and $\text{Var}(\varepsilon) = \text{diag}(\sigma^2\mathbf{I})$. Of course, $\mathbf{I}$ denotes

the $N \times N$ identity matrix. Motivated by the Giffen example from Sect. 2, we are interested in subgroups on which the parameter vector β significantly deviates from the parameter vector estimated on the whole dataset.

Given an estimate of the vector β, denoted $\hat{\beta}$, one can compute the vector of fitted values $\hat{Y}$. These quantities can be used to assess the appropriateness of the fitted model, by looking at the residuals $e = Y - \hat{Y}$. We estimate β with the ordinary least squares method, which minimizes the sum of squared residuals. This leads to the estimate:

$$\hat{\beta} = \left(\hat{\beta}_i\right) = \left(X^\top X\right)^{-1} X^\top Y$$

In order to define a proper quality measure for comparing estimated parameter vectors, we need to take into account the variance of the estimator $\hat{\beta}$, and the covariances between $\hat{\beta}_i$ and $\hat{\beta}_j$. For example, if $\hat{\beta}_i$ has a large variance compared to $\hat{\beta}_j$, then a given change in $\hat{\beta}_i$ should contribute less to the overall quality than the same change in $\hat{\beta}_j$, because the change in $\hat{\beta}_i$ is more likely to be caused by random variation. This suggest that

$$\left(\hat{\beta}^G - \hat{\beta}\right)^\top \left[\mathrm{Cov}\left(\hat{\beta}\right)\right]^{-1} \left(\hat{\beta}^G - \hat{\beta}\right)$$

might be a better distance measure than the normal Euclidean distance. In fact this expression is equivalent to Cook's distance up to a constant scale factor. R. Dennis Cook originally introduced his distance (Cook 1977) in 1977 for determining the contribution of single records to $\hat{\beta}$. He states that according to normal theory (Gentleman and Wilk 1975), the $(1-\alpha) \times 100\,\%$ confidence ellipsoid for the unknown vector, β, is given by the set of all vectors β^* satisfying

$$\frac{\left(\beta^* - \hat{\beta}\right)^\top \left[\widehat{\mathrm{Cov}}\left(\hat{\beta}\right)\right]^{-1} \left(\beta^* - \hat{\beta}\right)}{m} = \frac{\left(\beta^* - \hat{\beta}\right)^\top X^\top X \left(\beta^* - \hat{\beta}\right)}{ms^2} \leq F(m, N-m, 1-\alpha)$$

where

$$s^2 = \frac{e^\top e}{N-m} \qquad \widehat{\mathrm{Cov}}\left(\hat{\beta}\right) = s^2 \left(X^\top X\right)^{-1}$$

and $F(m, N-m, 1-\alpha)$ is the $1-\alpha$ probability point of the central F-distribution with m and $N-m$ degrees of freedom.

Now the stage has been set to determine the degree of influence of single records. Suppose we want to know how record r^i influences $\hat{\beta}$. Then one could naturally compute the least squares estimate for β with the record removed from the dataset. Let us denote this estimate by $\hat{\beta}_{(i)}$. We can adapt the confidence ellipsoid as an easily interpretable measure of the distance between $\hat{\beta}_{(i)}$ and $\hat{\beta}$. Hence, *Cook's distance* is defined as:

$$\Delta_i = \frac{\left(\hat{\beta}_{(i)} - \hat{\beta}\right)^\top X^\top X \left(\hat{\beta}_{(i)} - \hat{\beta}\right)}{ms^2} \tag{3}$$

Suppose for example that for a certain record r^i we find that $\Delta_i \approx F(m, N-m, 0.5)$. Then removing r^i moves the least squares estimate to the edge of the 50 % confidence region for β based on $\hat{\beta}$.

Cook and Weisberg extended Cook's distance to the case where multiple records are deleted simultaneously (Cook and Weisberg 1980). Let I be a vector of indices that specify the records to be deleted, and let $\hat{\beta}_{(I)}$ denote the least squares estimate for β computed from the dataset with all records in I removed. Cook's distance for multiple observations becomes

$$\Delta_I = \frac{\left(\hat{\beta}_{(I)} - \hat{\beta}\right)^\top X^\top X \left(\hat{\beta}_{(I)} - \hat{\beta}\right)}{ms^2} \tag{4}$$

and its geometric interpretation is identical to the geometrical interpretation of Δ_i. Any subset that has a large joint influence on the estimation of β corresponds to a large Δ_I.

For practical purposes one might not be interested in computing Cook's distance based on the entire parameter vector $\hat{\beta}$. For instance, one might be interested in the influence records have on the regression coefficient corresponding to one particular attribute, while excluding the intercept and other coefficients from the evaluation. To this end, Cook and Weisberg (1982) introduce the zero/one-matrix Z, with dimensions $m' \times m$, where m' is the number of elements of $\hat{\beta}$ that we are interested in (hence $m' \leq m$). The matrix Z is defined in such a way that $\psi = Z\beta$ are the coefficients of interest. Hence, if we are interested in the last m' elements of $\hat{\beta}$, Z starts from the left with $m - m'$ columns containing all zeroes, followed by a $m' \times m'$ identity matrix ($Z = (\mathbf{0}, \mathbf{I}_{\mathbf{m}'})$).

When using this transformation, Cook's distance (Eq. (4)) becomes:

$$\Delta_I^\psi = \frac{\left(\hat{\beta}_{(I)} - \hat{\beta}\right)^\top Z^\top \left(Z \left(X^\top X\right)^{-1} Z^\top\right)^{-1} Z \left(\hat{\beta}_{(I)} - \hat{\beta}\right)}{m's^2}$$

Since Cook's distance is invariant to changes in scale of the variables involved (Cook 1977), it would make an excellent quality measure for use in EMM:

Definition 6 (φ_{Cook}) Let D be a description. Its *quality according to Cook's distance* is given by:

$$\varphi_{\text{Cook}}(D) = \Delta_I^\psi, \text{ where } I = \left\{ i \middle| r^i \in \Omega, D\left(a_1^i, \ldots, a_k^i\right) = 0 \right\}$$

The quality of a description according to Cook's distance, is the distance bridged when the records that are not covered by the description are simultaneously discarded. Hence, Cook's distance is computed for the case where the records covered by the description D are *retained*.

6 Experimental results

For each EMM instance developed in the previous section, we run some experiments on relevant publicly available datasets in this section. EMM being an exploratory technique, our main evaluation method is interpreting the subgroups and fitted models using domain-specific literature. The reader should keep in mind that this type of analysis should normally be performed in collaboration with a subject area expert who could aid in the interpretation of the results. The experiments were run in Cortana (Meeng and Knobbe 2011), with some additional computations in R.[3] The publicly available version of Cortana contains implementations of the model classes Correlation (cf. Sect. 5.1) as target type DOUBLE_CORRELATION, Simple Linear Regression (cf. Sect. 5.3) as target type DOUBLE_REGRESSION, and Bayesian Network (cf. Sect. 5.5) as target type MULTI_LABEL. Recall that, in Sect. 4.1, we chose as description language $\mathcal{D}$ the conjunctions of conditions on single descriptors.

6.1 Correlation

For this model class we analyze the *Windsor housing* dataset[4] (Anglin and Gençay 1996). This dataset contains information on 546 houses that were sold in Windsor, Canada in the summer of 1987. The information for each house includes the two attributes of interest, $\ell_1 = x = lot\_size$ and $\ell_2 = y = sales\_price$. An additional 10 attributes are available to define candidate subgroups, including the number of bedrooms and bathrooms and whether the house is located at a desirable location. The correlation between lot size and sale price is 0.536, which implies that a larger size of the lot coincides with a higher sales price. The fitted regression function is:

$$y = 34136 + 6.60 \cdot x$$

As this function shows, on average one extra square meter corresponds to a sales price increase of $6.60.

On the Windsor housing dataset, we run an experiment with the significance of correlation difference measure, φ_{scd}. As discussed in Sect. 5.1, in order to be confident about the test results for the quality measure φ_{scd}, the support of a subgroup has to be over 25. This number was used as minimum support threshold for a run of Cortana using φ_{scd}. The following subgroup (and its complement) was found to show the most significant difference in correlation: $\varphi_{\text{scd}}(D_1) = 0.9993$.

$$D_1 : drive = 1 \wedge rec\_room = 1 \wedge nbath \geq 2.0$$

This is the group of 35 houses that have a driveway, a recreation room and at least two bathrooms. The scatter plots for the subgroup and its complement are given in Fig. 3. The subgroup shows a correlation of $\hat{r}^{G_1} = -0.090$ compared to $\hat{r}^{G_1^C} = 0.549$

[3] http://cran.r-project.org.

[4] Available from the Journal of Applied Econometrics Data Archive at http://econ.queensu.ca/jae/.

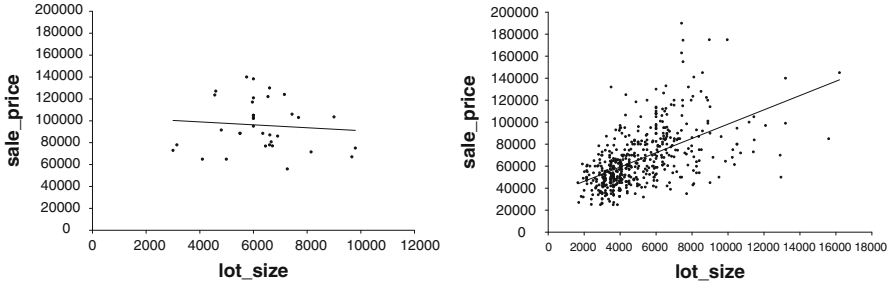

Fig. 3 Windsor housing, φ_{scd}: scatter plot of lot_size and $sales\_price$ for the subgroup corresponding to description $D_1 : drive = 1 \wedge rec\_room = 1 \wedge nbath \geq 2$ and its complement. **a** $G_1, \hat{r} = -0.090$, **b** $G_1^C, \hat{r} = 0.549$

Table 1 Subgroups on the housing data, and their sample correlation coefficients and supports

Subgroup	$\hat{r}$	n
Whole dataset	0.536	546
$nbath \geq 2$	0.564	144
$drive = 1$	0.502	469
$rec\_room = 1$	0.375	97
$nbath \geq 2 \wedge drive = 1$	0.509	128
$nbath \geq 2 \wedge rec\_room = 1$	0.129	38
$drive = 1 \wedge rec\_room = 1$	0.304	90
$nbath \geq 2 \wedge rec\_room = 1 \wedge \neg drive = 1$	−0.894	3
$nbath \geq 2 \wedge rec\_room = 1 \wedge drive = 1$	−0.090	35

for the remaining 511 houses. A tentative interpretation could be that G describes a collection of houses in the higher segments of the markets where the price of a house is mostly determined by its location and facilities. The desirable location may provide a natural limit on the lot size, such that this is not a factor in the pricing. Figure 3 supports this hypothesis: houses in G tend to have a higher price.

In general $sales\_price$ and lot_size are positively correlated, but EMM discovers a subgroup with a slightly negative correlation. However, the value in the subgroup is not significantly different from zero: a test of

$$H_0 : \hat{r}^{G_1} = 0 \quad \text{against} \quad H_1 : \hat{r}^{G_1} \neq 0$$

yields a p value of 0.61. The scatter plot confirms our impression that $sales\_price$ and lot_size are uncorrelated within the subgroup. For purposes of interpretation, it is interesting to perform some post-processing. In Table 1 we give an overview of the correlations within several subgroups whose intersection produces the final result, as given in the last row. It is interesting to see that the condition $nbath \geq 2$ in itself actually leads to a slight increase in correlation compared to the whole database, but the combination with the presence of a recreation room leads to a substantial drop to $\hat{r} = 0.129$. When we add the condition that the house should also have a driveway

Table 2 Data on gender and admission for various programs

n(g, a, p)	adm	Program		
Gender		A	B	C
Female	No	30	80	10
	Yes	10	80	40
Male	No	20	120	30
	Yes	40	120	20

Table 3 Quality (deviance) of several descriptions

D	$\varphi_{\text{dev}}(D)$
$Type = A$	21.09
$Type = C$	20.60
$Type = B$	0.28

we arrive at the final result with $\hat{r} = -0.090$. Note that adding this last condition only eliminates 3 records (the size of the subgroup goes from 38 to 35) and that the correlation between sales price and lot size in these three records (defined by the condition $nbath \geq 2 \wedge \neg drive = 1 \wedge rec\_room = 1$) is -0.894. We witness a phenomenon similar to Simpson's paradox: splitting up a subgroup with positive correlation (0.129) produces two subgroups both with a negative correlation (-0.090 and -0.894, respectively).

6.2 Association

As an example of the association model, suppose we are interested in the association between gender and admission to the master program, perhaps in an attempt to discover possible discrimination on gender. Suppose furthermore that one of the descriptive attributes is the type of master program (A, B, or C). The relevant artificial dataset is given in Table 2.

We can describe three subgroups by constraints on the attribute $type$; their quality is given in Table 3. For example, to evaluate the description $type = A$, we create a binary variable $type = A$ and then compare the fit of the homogeneous association model on $gender, admission, type = A$ against the saturated model. The homogeneous association model doesn't give a very good fit, resulting in a high deviance and thus high quality for the description. The association between gender and admission for master program A is quite different from the association between gender and admission for master programs B and C together.

To illustrate this EMM instance on a real data set, we consider a study on potential predictors for the completion or non-completion of a three-shot vaccine regimen (Chao et al. 2009). The data set contains 1413 observations on 10 variables, such as race, age, type of insurance, type of practice, location of the clinic, etc. Suppose we are interested in the relation between race and whether or not the person completes the regimen. To start with, let us consider the association between these two variables

Table 4 Association between race and completion of the vaccine regimen on the complete data set

Race	Completed?		P (Completed) (%)
	No	Yes	
White	452	280	38
Black	338	105	24
Hispanic	35	17	33
Other	119	67	36

Table 5 Association between race and completion of the vaccine regimen for subgroup $PracticeType = paediatric \wedge Age = 11\text{–}17\ years$ and its complement

Race	Subgroup			Complement		
	No	Yes	P (Completed) (%)	No	Yes	P (Completed) (%)
White	138	79	36	314	201	39
Black	139	41	23	199	64	24
Hispanic	20	8	29	15	9	38
Other	27	30	53	92	37	29

Columns labeled 'No' and 'Yes' indicate the number of persons who have not, or have, respectively, completed the regimen

in the complete data set (see Table 4). We summarize the association by giving the probability of completion for each race in the final column.

On a search depth of 2, the best subgroup that we find is $PracticeType = paediatric \wedge Age = 11\text{–}17\ years$, with a deviance of 10.59. The association between race and completion for this subgroup is given in Table 5. The most striking difference is that in the subgroup the probability of completion of the regimen for Race = Other is much higher than in the overall data base (and the complement of the subgroup). In the subgroup the probability of completion is 53 %, whereas in the complement group it is only 29 %.

6.3 Simple linear regression

On the Windsor housing dataset using the Significance of Slope Difference (φ_{ssd}) quality measure, we find as highest ranking subgroup the 226 houses that have a driveway, no basement and at most one bathroom:

$$D_2 : drive = 1 \wedge basement = 0 \wedge nbath \leq 1$$

The subgroup G_2 and its complement G_2^C (320 houses) lead to the following two fitted regression functions, respectively:

$$G_2 : \quad y = 41568 + 3.31 \cdot x$$
$$G_2^C : \quad y = 30723 + 8.45 \cdot x$$

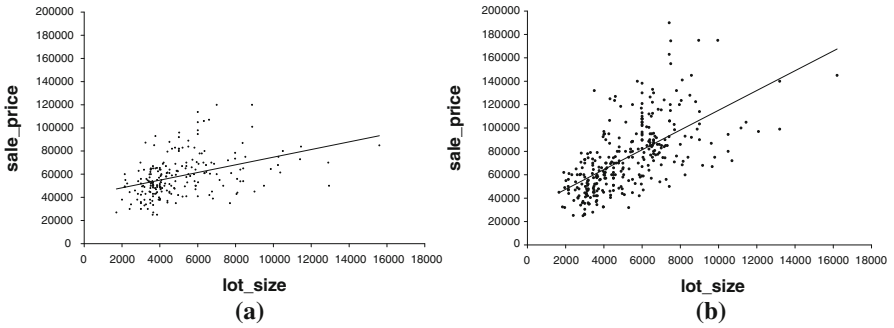

Fig. 4 Windsor housing, φ_{ssd}: scatter plot of lot_size and $sales\_price$ for the subgroup corresponding to $D_2 : drive = 1 \wedge basement = 0 \wedge nbath \leq 1$ and its complement. **a** G_2, $y = 41568 + 3.31 \cdot x$, **b** G_2^C, $y = 30723 + 8.45 \cdot x$

The subgroup quality is $\varphi_{\text{ssd}} > 0.9999$, meaning that the p value of the test

$$H_0 : \beta_1^{G_2} = \beta_1^{G_2^C} \quad \text{against} \quad H_1 : \beta_1^{G_2} \neq \beta_1^{G_2^C}$$

is virtually zero. As discussed in Sect. 3.2.1, this does not necessarily imply that the subgroup is statistically significant; a low p value merely implies that the subgroup is more exceptional than other subgroup with higher p values, but no significance claim can be made due to the multiple comparisons problem (cf. Sect. 3.2.1). There are subgroups with a larger difference in slope, but the reported subgroup scores higher because it is quite big. Figure 4 shows the scatter plots of lot_size and $sales\_price$ for the subgroup and its complement.

6.4 Classification

We demonstrate the classification model in the domain of bioinformatics, on the *Affymetrix* dataset. In genetics, genes are organized in so-called *gene regulatory networks*. This means that the expression (its effective activity) of a gene may be influenced by the expression of other genes. Hence, if one gene is regulated by another, one can expect a linear correlation between the associated expression-levels. In many diseases, specifically cancer, this interaction between genes may be disturbed. The Gene Expression dataset shows the expression-levels of 313 genes as measured by an Affymetrix microarray, for 63 patients that suffer from a cancer known as neuroblastoma (van de Koppel et al. 2007). Additionally, the dataset contains clinical information about the patients, including age, sex, stage of the disease, etc.

In the logistic regression experiment, we take *NBstatus* as the output ℓ_2=y, and age (age at diagnosis in days) as the predictor $\ell_1 = x$. The subgroups are created using the gene expression level variables. Hence, the model specification is

$$\begin{aligned} &\text{logit}\{P(NBstatus = \text{`relapse or deceased'})\} \\ &= \beta_0 + \beta_1 \cdot D(i) + \beta_2 \cdot x + \beta_3 \cdot (D(i) \cdot x) \end{aligned}$$

We find the subgroup

$$D_3 : \text{SMPD1} \geq 840 \wedge \text{HOXB6} \leq 370.75$$

with a coverage of 33, and quality $\varphi_{\text{sed}}(D_3) = 0.994$. We find a positive coefficient of x for the subgroup, and a slightly negative coefficient for its complement. Within the subgroup, the odds of *NBstatus* = 'relapse or deceased' increase with 44 % when the age at diagnosis increases with 100 days, whereas in the complement the odds decrease with 8 %. More loosely, within the subgroup an increase in age at diagnosis decreases the probability of survival, whereas in the complement an increase in age slightly increases the probability of survival. Such reversals of the direction of influence may be of particular interest to the domain expert.

6.5 Bayesian network

So far, we have discussed results on model classes with a severely restricted number of targets. This has the benefit that multiple facets of the resulting subgroups can be interpreted: we can interpret the description of the subgroup, and we can also interpret the associated model. While this model interpretation can give us valuable information on the found subgroups, EMM was designed to capture intricate interactions between a multitude of targets. Hence, restricting our attention to model classes with only two or three targets, implies restricting the expressive power of EMM.

In this section and in Sect. 6.6, we discuss results on model classes that allow the number of targets to grow as large as the user wants (or the data allows). Since the resulting subgroups capture unusual interplay between these potentially numerous targets, the model classes have a large expressive power. The obvious drawback to this power is that the associated models are so intricate that they become impossible to interpret completely by the human eye. One can still examine some cherrypicked particular elements of the resulting models, but the overview one has with the simpler model classes is gone. Relatively complex model classes ask the Exceptional Model Miner to suspend their disbelief on the target space: since the miner has determined which quality measure is used to find exceptional models, he/she will have to trust that the found models are indeed exceptional. The subgroups can of course still be interpreted on the descriptor space: the complexity of the descriptions doesn't change with the chosen model class.

6.5.1 Emotions data

The *Emotions* dataset (Trohidis et al. 2008) 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*.

We obtain the networks shown in Fig. 5. Figure 5a depicts a network fitted on the whole dataset, and Fig. 5b displays a network fitted on a subgroup of size 94 corresponding to description D_4 : *STD_MFCC_7* $\leq$ 0.203 and *Mean_Centroid* $\geq$

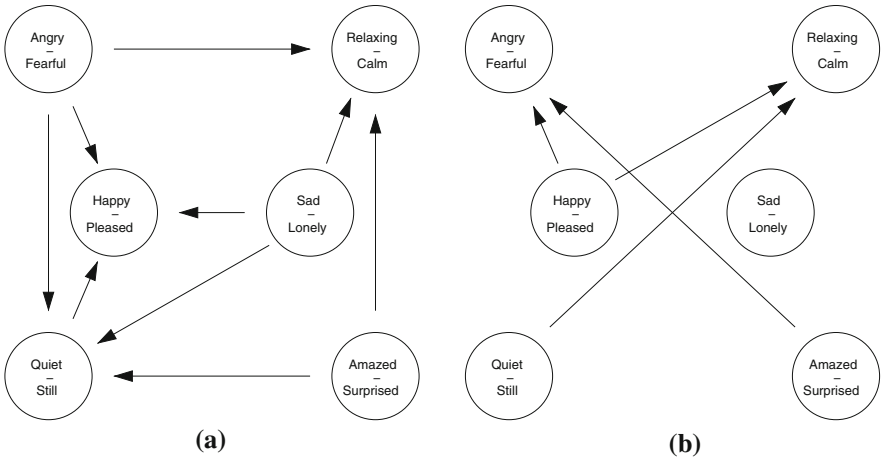

Fig. 5 Bayesian networks for the *Emotions* data. **a** Whole dataset, **b** D_4 : *STD MFCC 7* ≤ 0.203 and *Mean Centroid* ≥ 0.066

0.066, with quality $\varphi_{\text{weed}}(D_4) = 0.675$. The first condition says that coefficient 7 of the 13-band Mel Frequency Cepstrum has a 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 (Schubert et al. 2004).

From Fig. 5a 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. Fig. 5b). By lack of experts on the domain of this dataset, we will refrain from interpreting the Bayesian networks, also keeping in mind the discussion in the second paragraph of Sect. 6.5.

6.5.2 Mammals data

The *Mammals* dataset (Garriga et al. 2007; Mitchell-Jones et al. 1999) 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 details of the corresponding land areas. We define candidate subgroups by conditions on the climate data, and fit Bayesian networks on the mammals. We use a version of this dataset that was pre-processed by Heikinheimo et al. (2007).

The found exceptional subgroup G_5 is defined by the constraints *max_temp_mar* ≤ 7.97 and *max_temp_sep* ≤ 17.65, i.e. the temperatures in both March and September do not reach high levels. It has quality $\varphi_{\text{weed}}(D_5) = 0.121$, and size $|G_5| = 834$. Existing studies of the Mammals dataset provide evidence that the discoveries corre-

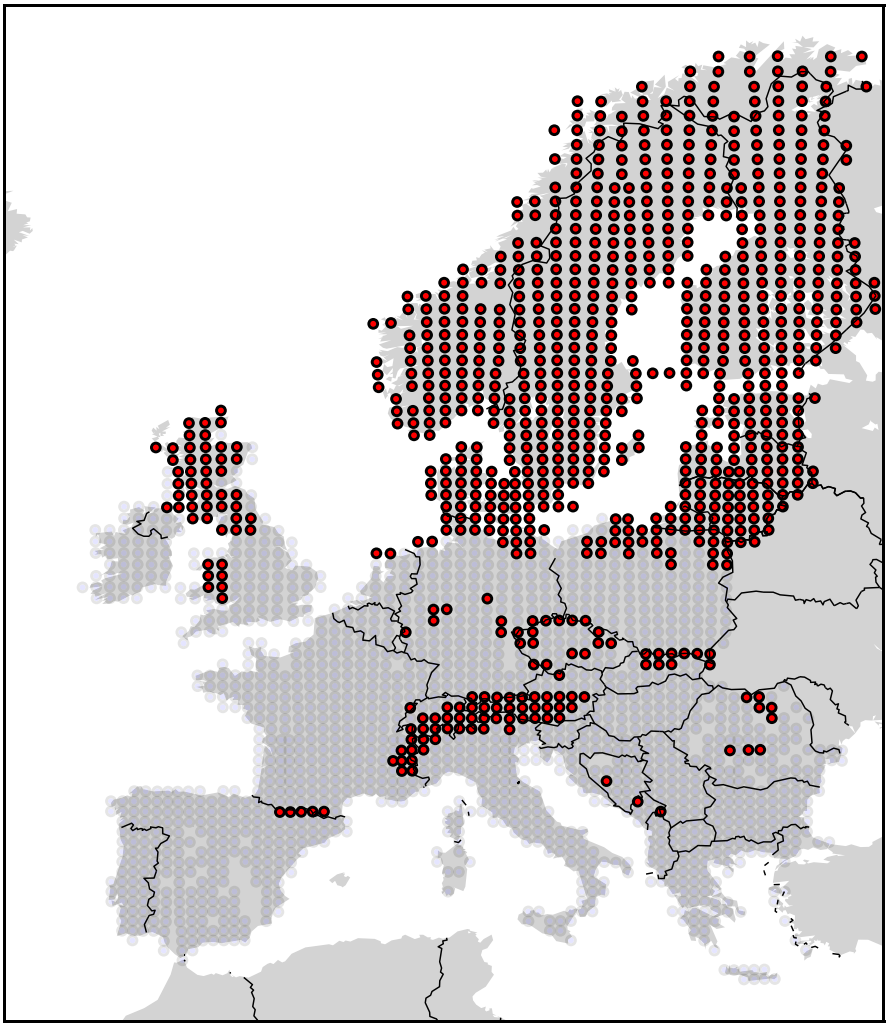

Fig. 6 Regions in Europe that belong to the subgroup defined as D_5 : *max temp mar* $\leq$ 7.97 and *max temp sep* $\leq$ 17.65 ($n = 834$)

spond to real underlying concepts in the dataset domain, by considering the spatial coherence of the discovery on the map of Europe (Heikinheimo et al. 2007). Mirroring this evaluation technique, Fig. 6 shows the regions in Europe that belong to the subgroup. At first sight, the subgroup seems to consist of relatively coherent Northern parts of Europe, and some random clutter more to the South. However, when one overlays a map of Europe that depicts the altitudes, this clutter turns out to be much more coherent than it seems: these are the not-too-Southern mountainous regions of Europe, including the Alps, the Pyrenees, and the Carpathians. These observations do not necessarily *prove* that the subgroup describes an underlying concept in the dataset, but they do provide evidence.

Summing up all conditional dependence relations in a 101-node Bayesian network would results in a list that is unsurveyable to the average reader, so instead, we highlight two particular relations that differ between the whole dataset and the subgroup. On the whole dataset, but not on G_5, the European Water Vole (*Arvicola terrestris*) and the Mountain Hare (*Lepus timidus*) are conditionally dependent given the Ermelin (*Mustela erminea*). Conversely, on G_5, but not on the whole dataset, the Red Squirrel (*Sciurus vulgaris*) and the Least Weasel (*Mustela nivalis*) are conditionally dependent given the European Badger (*Meles meles*).

6.6 General linear regression

6.6.1 Giffen behavior data

This dataset was used for a study that provided the first real-world evidence of Giffen behavior, i.e. an upward sloping demand curve (Jensen and Miller 2008). As common sense suggests, the demand for a product will usually decrease as its price increases. According to economic textbooks, there are circumstances however, for which we should expect to see an upward sloping demand curve. The common example is that of poor families that spend most of their income on a relatively inexpensive staple food (e.g. rice or wheat) and a small part on a more expensive type of food (e.g. meat). If the price of the staple food rises, people can no longer afford to supplement their diet with the more expensive food, and must consume more of the staple food.

The dataset we analyze was collected in different counties in the Chinese province Hunan, where rice is the staple food. The price changes were brought about by giving vouchers to randomly selected households to subsidize their purchase of rice. For each household, two changes are observed: the change between periods 2 and 1 ($t = 2$), capturing the effect of giving the subsidy; and the change between periods 3 and 2 ($t = 3$) capturing the effect of removing the subsidy. The global model estimated in Jensen and Miller (2008) is:

$$\Delta\text{staple}_{i,t} = \beta_0 + \beta_1 \Delta p_{i,t} + \sum \beta_2 \Delta Z_{i,t} + \sum \beta_3 \text{County} \times \text{Time}_{i,t} + \Delta\varepsilon_{i,t}$$

where $\Delta\text{staple}_{i,t}$ denotes the percent change in household i's consumption of rice, $\Delta p_{i,t}$ is the percent change in the price of rice due to the subsidy (negative for $t = 2$ and positive for $t = 3$), and $\Delta Z_{i,t}$ is a vector of percent changes in other control variables including income and household size. County $\times$ Time denotes a set of dummy variables included to control for any county-level factors that change over time. For further details about the design of the study and the estimation strategy, we refer to Jensen and Miller (2008).

The coefficient of primary interest is β_1. If $\beta_1 > 0$ we observe Giffen behavior. The other variables are included in the model to control for other possible influences on demand, so that the effect of price can be reliably estimated. Therefore it makes sense to restrict our quality measure to the coefficient β_1.

Jensen and Miller (2008) suggest that for the extremely poor, one might not observe Giffen behavior, because they consumed rice almost exclusively anyway, and therefore

have no other possibility than buying less of it in case of a price increase. The Initial Staple Calorie Share (ISCS) was also measured in the study, and the hypothesis is that families with a high value for this variable do not display Giffen behavior. (Jensen and Miller 2008) tried several manually selected thresholds on ISCS; for example, for the subgroup of households with ISCS > 0.8, indeed it is observed that $\hat{\beta}_1 = -0.585$ (no Giffen behavior) whereas for ISCS ≤ 0.8 they get $\hat{\beta}_1 = 0.466$ (Giffen behavior).

We analyzed this dataset with ISCS as one of the variables on which the subgroups could be defined. The best subgroup we found was ISCS ≥ 0.87 with $\hat{\beta}_1 = -0.96$ (against $\hat{\beta}_1 = 0.22$ for the complete dataset). The size of this subgroup is $n = 106$. This confirms the conclusion that Giffen behavior does not occur for families that almost exclusively consume rice anyway. This conclusion can also be reached by defining subgroups on *income per capita* rather than ISCS. Particularly illustrative examples are the 4th-ranked subgroup: Income per Capita ≤ 64.67, with a slope of -0.41, and the 6th-ranked subgroup: Income per capita ≥ 803.75, with a slope of 0.79 (strong Giffen behavior).

6.6.2 EAEF data

This dataset was extracted from the National Longitudinal Survey of Youth 1979- (NLSY79). It contains information about hourly earnings of men and women, their education, and other information. For more details, we refer to Appendix B of Dougherty (2011). We fit a model relating years of schooling and years of work experience to earnings. The model fitted on the complete dataset is:

$$\text{Earnings} = -29.15 + 2.78 \times \text{YrsOfSchool} + 0.63 \times \text{YrsWorkExp}$$

All coefficients in this model are significant at $\alpha = 0.01$, and R^2 is approximately equal to 20 %.

The 4$^{\text{th}}$ ranked subgroup we found was COLLBARG = 1, meaning that the pay was set by collective bargaining. The fitted model for this subgroup of size $n = 533$ is:

$$\text{Earnings} = -8.93 + 1.57 \times \text{YrsOfSchool} + 0.43 \times \text{YrsWorkExp}$$

This suggests that for this group an extra year of schooling on average leads to an increase of just \$1.57 in hourly earnings, compared to \$2.78 for the whole dataset. The same is true for the influence of an extra year of work experience: just \$0.43 for the collective bargaining subgroup, against \$0.63 in the complete dataset. This is consistent with the finding that unions tend to equalize the income distribution, especially between skilled and unskilled workers (Aidt and Tzannatos 2002).

6.6.3 Personal computer data

This dataset was analyzed in Stengos and Zacharias (2006). The data was collected from advertisements in PC Magazine. Each observation consists of the advertised

price and features of personal computers. We have fitted the following model to the complete dataset:

$$\begin{aligned}\text{Price} = &-246.68 + 8.89 \times \text{Spd} + 0.71 \times \text{HD} + 47.39 \times \text{RAM} \\ &+ 126.70 \times \text{Scr} + 0.97 \times \text{Ads} - 47.08 \times \text{Trend}\end{aligned}$$

where *Price* is the price in US dollars of a 486 PC, *Spd* is the clock speed in MHz, *HD* is the size of the hard drive in MB, *RAM* is the size of RAM in MB, *Scr* is the size of the screen in inches, *Ads* is the number of 486 price listings in the month the advertisement was placed, and *Trend* is a time trend indicating month starting from January of 1993 (Trend = 1) to November of 1995 (Trend = 35). All coefficient estimates are significant at $\alpha = 0.01$, and R^2 is about 71 %.

By far the most important attribute to create subgroups was whether or not the company was a "premium firm" (IBM or COMPAQ). The most deviating subgroup were the non-premium firms:

$$\begin{aligned}\text{Price} = &-2130.21 + 13.15 \times \text{Spd} + 2.31 \times \text{HD} + 22.20 \times \text{RAM} \\ &+ 252.80 \times \text{Scr} + 0.79 \times \text{Ads} - 46.45 \times \text{Trend}\end{aligned}$$

The size of this subgroup is $n = 612$, and R^2 is about 85 %. We get the clearest picture when we contrast this with the regression model fitted to the premium firms:

$$\begin{aligned}\text{Price} = &\ 165.69 + 8.50 \times \text{Spd} + 0.67 \times \text{HD} + 53.66 \times \text{RAM} \\ &+ 99.96 \times \text{Scr} + 0.65 \times \text{Ads} - 47.87 \times \text{Trend}\end{aligned}$$

The size of this subgroup is $n = 5647$, and R^2 is about 79 %. We find mostly reasonable behavior in these subgroups: the price of computers from premium firms is based on a far higher intercept, since the premium brand name ensures a vast price upkeep. Consequently, other factors have a substantially smaller impact on the price than for computers from non-premium firms. Oddly, the size of RAM memory does matter more strongly for premium brands than for non-premium brands.

6.7 Experiments with alternative quality measures

To illustrate how one can influence the results of EMM by the selection of quality measure, we explore some results found with alternative measures for a few model classes.

6.7.1 Correlation

For the Correlation model class, we run additional experiments on the *Affymetrix* dataset (cf. Sect. 6.4) with the alternative quality measure φ_{abs} (cf. Sect. 5.1.1). Recall that this measure computes the absolute difference of the correlation for a description

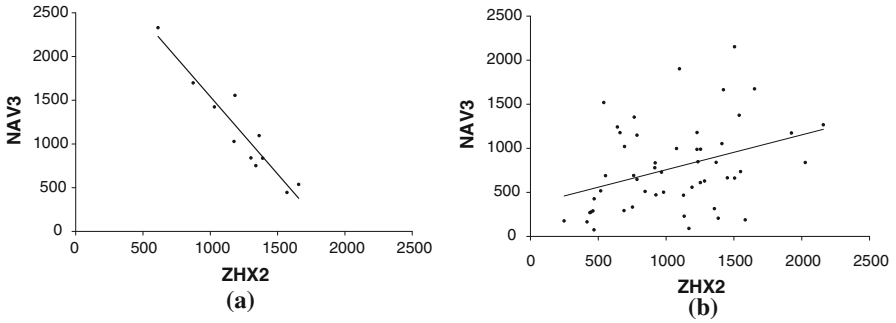

Fig. 7 Affymetrix, φ_{abs}: scatter plot of the subgroup corresponding to description D_6 : $11\_band = \text{`}nodeletion\text{'} \wedge survival\ time \leq 1919 \wedge XP\_498569.1 \leq 57$ and its complement. **a** G_6, with $\hat{r} = -0.950$, **b** G_6^C, with $\hat{r} = 0.363$

and its complement, without considering the coverage of the descriptions, and hence does not do anything to prevent overfitting.

On the *Affymetrix* dataset, recall that we analyze the correlation between *ZHX3* and *NAV3*, showing a very slight correlation ($\hat{r} = 0.218$) on the whole dataset. We analyze this dataset in terms of the absolute difference of correlations φ_{abs}, allowing the use of all remaining attributes (both gene expression and clinical information) for building descriptions. As the φ_{abs} measure does not have any provisions for promoting larger subgroups, we use a minimum support threshold of 10 (15 % of the patients). The largest distance ($\varphi_{abs}(D_6) = 1.313$) was found with the following description covering 11 records (17.5 %) of the dataset

$$D_6 : 11\_band = \text{`}no\ deletion\text{'} \wedge survival\ time \leq 1919 \wedge XP\_498569.1 \leq 57$$

Figure 7 shows the plot for this description and its complement with the regression lines drawn in. The correlation for the description is $\hat{r}^{G_6} = -0.95$ and the correlation in the remaining data is $\hat{r}^{G_6^C} = 0.363$. Note that the description displays a very "stable" behavior: all points are quite close to the regression line, with $R^2 \approx 0.9$.

6.7.2 Classification

As an alternative in the Classification model class, we run experiments with a DTM classifier on the *Affymetrix* dataset, evaluating subgroups with both alternative quality measures (cf. Sect. 5.4.1). We have selected three binary attributes as targets. The first two attributes, which serve as input variables of the decision table, are related to genomic alterations that may be observed within the tumor tissues. The attribute *1p_band* (ℓ_1) describes whether the small arm ('p') of the first chromosome has been deleted. The second attribute, *mycn* (ℓ_2), describes whether one specific gene is amplified or not (multiple copies introduced in the genome). Both attributes are known to potentially influence the genesis and prognosis of neuroblastoma. The output attribute for the classification model is *NBstatus* (ℓ_3), which can be either *'no event'* or *'relapse*

or deceased'. The following decision table describes the conditional distribution of *NBstatus* given *1p_band* and *mycn* on the whole dataset

1p_band =	*mycn* = *'amplified'*	*mycn* = *'not amplified'*
'deletion'	0.333 (3)	0.667 (3)
'no change'	0.625 (8)	0.204 (49)

In order to find descriptions for which the distribution is significantly different, we run EMM with the Hellinger distance φ_{Hel} as quality measure. As our quality measures for classification do not specifically promote descriptions with larger coverage, we have selected a slightly higher minimum support threshold of 16, which corresponds to 25 % of the data. The following subgroup of 17 patients (27.0 %) was the best found ($\varphi_{\text{Hel}}(D_7) = 3.803$)

$$D_7 : prognosis = \text{`}unknown\text{'}$$

1p_band =	*mycn* = *'amplified'*	*mycn* = *'not amplified'*
'deletion'	1.0 (1)	0.833 (6)
'no change'	1.0 (1)	0.333 (9)

Note that for each combination of input values, the probability of *'relapse or deceased'* is increased, which makes sense when the prognosis is uncertain. Note furthermore that the overall dataset does not yield a pure classifier: for every combination of input values, there is still some confusion in the predictions.

In our second alternative classification experiment, we are interested in "predictable" descriptions. Therefore, we run EMM with the φ_{BDeu} measure. All other settings are kept the same. The following subgroup ($|G_8| = 16$ (25.4 %), $\varphi_{\text{BDeu}}(D_8) = -1.075$) is based on the expression of the gene *RIF1* ('RAP1 interacting factor homolog (yeast)')

$$D_8 : RIF1 \geq 160.45$$

1p_band =	*mycn* = *'amplified'*	*mycn* = *'not amplified'*
'deletion'	0.0 (0)	0.0 (0)
'no change'	0.0 (0)	0.0 (16)

For this description, the predictiveness is optimal, as all patients turn out to be tumor-free. In fact, the decision table ends up being rather trivial, as all cells indicate the same decision.

6.7.3 Bayesian network

In Sect. 5.5, 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_9 : *mean_temp_apr* $\geq 11.86 \wedge$ *mean_temp_aug* ≤ 23.28. Its quality is $\varphi_{\text{ed}}(D_9) = 0.147$, and its coverage is $|G_9| = 105$ (4.7 %). The regions in Europe that belong to D_9 are displayed in Figure 8.

Relations between mammals that distinguish D_9 from Ω (i.e.: relations that appear as v-structures in the one Bayesian network but not in the other) include the following. On Ω, but not on D_9, 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_9, 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_9 found with φ_{ed}, with the description D_5 found with φ_{weed}, there are several things to remark. As expected, using the plain edit distance leads EMM to report 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 have discussed in Sect. 6.5.2, the area depicted in Fig. 6 seems to indicate that subgroup G_5 spans a dichotomous but relatively coherent part of Europe: some Northern areas, and some mountainous areas. By contrast, the regions belonging to subgroup G_9, as depicted in Fig. 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,

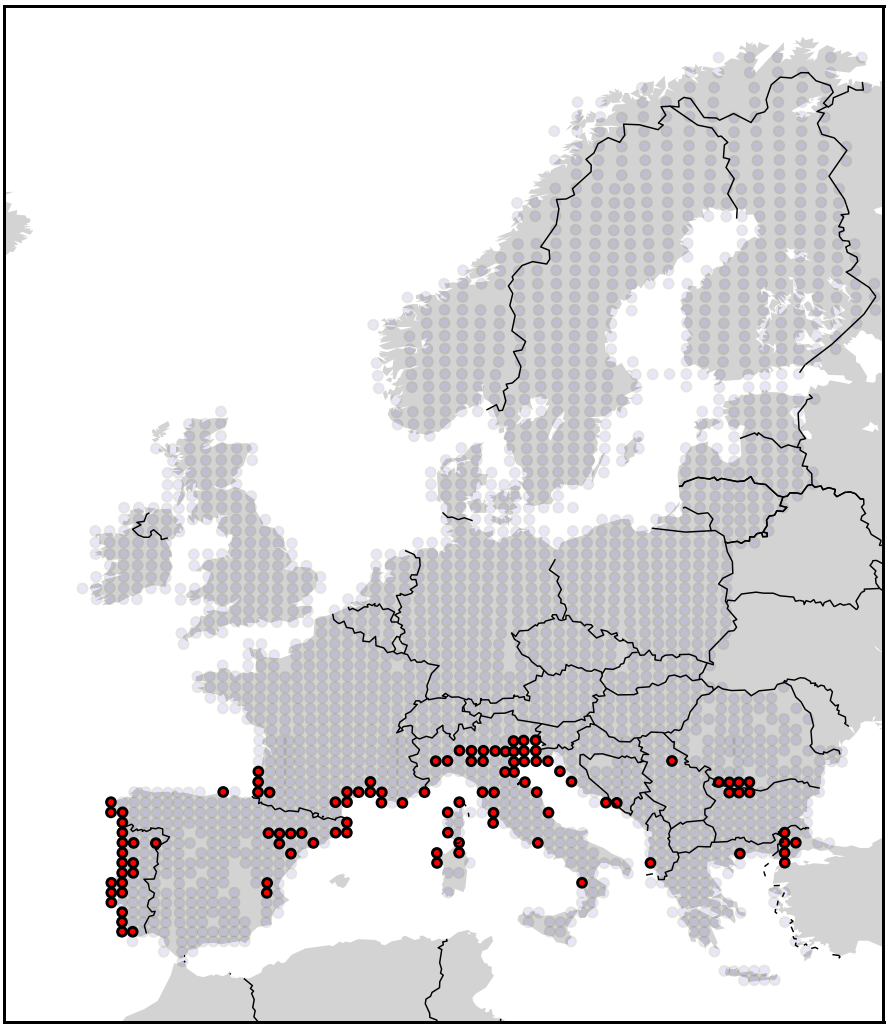

Fig. 8 Regions in Europe that belong to the subgroup corresponding to D_9 : *mean_temp_apr* $\geq 11.86 \wedge$ *mean_temp_aug* ≤ 23.28 ($|G_9| = 105$)

roughly speaking, φ_{weed} seems to deliver more substantially coherent subgroups than φ_{ed}.

7 Related work

EMM extends a vast body of work, of which this section contains some highlights. First, we discuss the search strategies developed to deal with the exponential search space. Then, we look into other local pattern mining tasks, and other extensions of SD. Finally, we discuss how similar questions arise in other data mining disciplines, and what distinguishes them from EMM.

7.1 Search strategies for SD/EMM

When striving to find interesting subsets of a dataset, the search space is exponential in the number of records. By restricting the problem to finding interesting *subgroups*, i.e. subsets with a concise description, the search space remains theoretically exponential (when the description language is generous, and the attributes are high-cardinality numerical or nominal), but we obtain a handle with which we can tackle the problem. Traditionally (Klösgen 1996), this is done by compelling all attributes in the dataset to be nominal (and assuming their cardinality to be not too large). In this case, exhaustive search is possible, using filters akin to the anti-monotonicity constraints known from frequent itemset mining. An interesting approach employing optimistic estimates for a continuous target is given in Atzmüller and Lemmerich (2009), but here too, all descriptors are compelled to be nominal. The same holds for the Non-redundant Subgroup Discovery task (Boley and Grosskreutz 2009), where the goal is to deliver only those subgroups that form a representative for an equivalence class with respect to their extensions in the dataset: though the approach is undeniably a valuable contribution to the field, all attributes are compelled to be nominal. When not all attributes are nominal, traditionally there was no option other than to resort to heuristic search.

Instead of opting for heuristic search, one could of course make the numeric attributes nominal by static discretization before invoking the SD algorithm. The obvious drawback is that by doing this, one loses information on the numeric attribute. A refreshing alternative to discretization which ameliorates this drawback, is by employing fuzzy partitions in an evolutionary algorithm (del Jesús et al. 2007). The numeric attributes are still essentially discretized, but during instead of before the algorithm, and in a flexible, reversible manner. This way, we can expect to retrieve more information from the attributes than static discretization would allow. This fuzzy innovation is combined with mining non-dominated subgroups in terms of support and unusualness in Carmona et al. (2010).

Recently, Grosskreutz and Rüping (2009) developed a new pruning scheme with accompanying SD algorithm, MergeSD, which allows for exhaustive mining even when the attributes are taken from a numerical domain. Their key idea is to exploit bounds between related numerical subgroup descriptions to prune with optimistic estimates, thus reducing the search space to tractable levels. Unfortunately, the pruning scheme cannot be used with any quality measure; implicitly a constraint similar to anti-monotonicity is imposed, bounding the choice of target concept and quality measure and hence restricting the pruning scheme to the subset of EMM instances for which anti-monotonicity can be enforced.

In work dedicated to expanding the description language available to Subgroup Discoverers, Mampaey et al. introduced an efficient treatment of numerical attributes (Mampaey et al. 2012). The subgroup space is not explored exhaustively. Instead, the algorithm finds richer descriptions efficiently, by finding an optimal interval for every numerical attribute, and an optimal value set for every nominal attribute. The efficiency comes from having the algorithm only consider subgroup descriptions that lie on a convex hull in ROC space, and evaluating subgroups with a convex quality measure. Hence, the method is only suitable for a target concept that can be properly

expressed in ROC space, i.e. traditional SD with a nominal target, and a convex concept of interestingness.

Another problem stemming from the exponential search space, is the redundancy in a resulting subgroup set. When a subgroup is deemed interesting, it is very likely that small tweaks to the subgroup will lead to other subgroups that are also quite interesting. Therefore it is not uncommon, especially when there are numerical attributes in the dataset, to find the top of a subgroup chart being dominated by many copies of what technically may all be slightly different subgroups, which in practice all indicate the same underlying concept. Scholz (2005) binds the utility of a subgroup to prior knowledge, which encompasses all previously found subgroups. Iterative modeling evaluates candidate subgroups by weighing data according to the prior knowledge. The data distribution is updated in each iteration, such that redundant patterns no longer receive a high utility. Lavrač et al. (2004) employ a beam search approach which is very similar to the one in Algorithm 1 in the current paper, except for two properties: their beam search uses the classification accuracy of a rule as quality measure, and they select only a single rule through the beam search. The deployment of a weighted covering scheme before re-running the beam search algorithm, leads to a rule set with increased diversity. van Leeuwen and Knobbe (2011) introduced three degrees of subgroup redundancy, and incorporated selection strategies based on these redundancies in a beam search algorithm. This results in non-exhaustive, but interestingly different search strategies. The authors extended their approach in a paper (van Leeuwen and Knobbe 2012) where some subgroup quality is conceded to increase the diversity of the result set. This is achieved by post-processing the results of a larger beam search run.

Whereas the traditional EMM framework strives to find exceptional subgroups by searching through the descriptive attribute space, and evaluating on the target attribute space, interesting results have been obtained by taking a more symmetrical approach to the two subspaces of the data. The EMDM algorithm (van Leeuwen 2010) strives to effectively find exceptional models by iteratively improving candidate subgroups, exploiting structure in both spaces. Each iteration consists of two steps, one for Exception Maximization (EM) and one for Description Minimization (DM). In the EM step, a compression-based quality measure guides the search for subsets having an unusual model. In the DM step, a rule-based classifier is employed to find a concise description that crafts a subgroup from the found subset. Upon convergence, or when a threshold on the number of iterations is surpassed, the subgroups are reported.

New developments have also been made towards exhaustive EMM, by adapting the well-known FP-Growth algorithm. The *generic pattern growth* algorithm (GP-Growth) of Lemmerich et al. (2012) strives to avoid scanning the whole dataset to evaluate subgroups. Instead, it builds a special datastructure, in which the key information of the model learned for a subgroup is summarized. Such a summary is called a *valuation basis*. It contains enough information to determine the quality of any refinement of the subgroup. The GP-Growth algorithm can reduce the memory requirement and runtime of an EMM instance by more than an order of magnitude, but only when a valuation basis can be found that is suitably condensed. This depends on the chosen model class: for relatively simple model classes it can be done, but for the more computationally expensive model classes it cannot. If a parallel single-pass algorithm

with sublinear memory requirements exists to compute the model from a given set of records, profit can be gained from GP-Growth. Most of the model classes discussed in Sect. 5 can benefit from GP-Growth, but the Bayesian Network model class of Sect. 5.5 cannot.

7.2 Similar local pattern mining tasks

SD research originated in the mid-nineties, in a simple single-table setting with a binary target attribute (Klösgen 1996), and in a multi-relational setting (Wrobel 1997). The latter paper has a very general definition of an *evaluation function*, which can be seen as equivalent to Definition 2 from this paper. At the same time, several papers were written giving a different name to essentially the same task, most notably Bump Hunting (Friedman and Fisher 1999) and Data Surveying (Siebes 1995).

At the turn of the century, Willi Klösgen wrote a pair of survey papers on SD (Klösgen 1998, 1999). The first (Klösgen 1998) discusses interestingness, description languages, subgroup validation, search strategies, and presentation of the results. Particularly interesting is that this paper looks beyond the standard binary, nominal, and numeric attributes, by investigating SD on time-stamped, spatial, and text-based data. The second (Klösgen 1999) again explores interestingness, and adds a discussion on conditions that make SD successful, as well as a few applications.

Tasks that are very similar to, but slightly different from Subgroup Discovery, include Contrast Set Mining (Bay and Pazzani 2001), where the goal is to find "conjunctions of attributes and values that differ meaningfully in their distributions across groups", and Emerging Pattern Mining (Dong and Li 1999), which strives to find itemsets whose support increases substantially from one dataset to another. Kralj Novak et al. (2009) provide a framework unifying Contrast Set Mining, Emerging Pattern Mining, and Subgroup Discovery.

For nominal-valued datasets, Zimmermann and Raedt (2009) have introduced the problem of Cluster-Grouping as a bridge between Subgroup Discovery and several other important data mining tasks, such as clustering and classification.

Giving a full overview of all work related to SD is beyond the scope of this paper; such overviews are available in the literature (for instance: (Herrera et al. 2011)). In the remainder of this section, we focus on work related to supervised local pattern mining with a more complex goal in mind.

As the antithesis to Contrast Set Mining, Redescription Mining (Ramakrishnan et al. 1995; Gallo et al. 2008) strives to find multiple descriptions of the same subgroups, originally in itemset data. It has recently been extended to categorical and real-valued data (Galbrun and Miettinen 2012).

Umek and Zupan (2011) consider SD with a multi-dimensional output space. They approach this data by considering the output space first: by agglomerative clustering in the output space, candidate subgroups are proposed that have records similar in outcomes. Then, a predictive modeling technique is used to test for each identified candidate whether it can be characterized by a description over the input space.

One of the few papers that explicitly seeks a deviating model over a target attribute, concerns Distribution Rules (Jorge et al. 2006). In this work, there is only one numeric

target, and the goal is to find subgroups for which the distribution over this target is significantly different from the overall distribution, measured in terms of the Kolmogorov-Smirnov test for goodness of fit. Since rules are evaluated by assessing characteristics of a model, this can be seen as an early instance of EMM, albeit considering only one target attribute.

An extensive overview of the state of the art in Local Pattern Mining (LPM) was portrayed in a Dagstuhl Seminar in 2004. Its proceedings (Morik et al. 2005) provide a wealth of approaches to LPM by experts from many data mining and machine learning subfields that are not necessarily primarily focused on Local Pattern Mining. Hence, the proceedings provide many interesting alternative approaches.

7.3 Similar tasks with a broader scope

General concepts from EMM, like fitting different models to different parts of the data, or identifying anomalies in a dataset, appear in tasks beyond Local Pattern Mining. In this section we discuss a few such tasks, and how they relate to EMM.

In Outlier Detection, traditionally the goal is to identify records that deviate from a general mechanism. Usually there is no desire to find a coherent set of such outliers, which can succinctly be described: identifying non-conforming records is enough. Early on, Hand et al. (2002) pointed at the distinction between noise, i.e. exceptions to be deleted, and local patterns, which form interesting subgroups. As Outlier Detection becomes more and more mature and sophisticated, we witness more attention towards the reason why a point is an outlier, for instance in recent work by Kriegel et al. (2012). Their method to detect outliers in arbitrarily oriented subspaces of the original attribute space also delivers an explanation with each outlier, consisting of two parts: an error vector, pointing towards the expected position of the outlier, and an outlier score, indicating the likelihood that the outlier is generated by a different mechanism rather than being just a rare object from the general mechanism. Searching for the reason for outliers is a step towards bridging the gap with finding coherent deviating subsets as done in EMM, although the approaches differ vastly.

When fitting a regression function to a dataset with a complex underlying distribution, one could employ Regression Clustering (Zhang 2003). The idea is to simultaneously apply $K > 1$ regression functions to the dataset, clustering the dataset into K subsets that each have a simpler distribution than the overall distribution. Each function is then regressed to its own subset, resulting in smaller residual errors, and the regression functions and clustering optimize a common objective function. Catering for parts of the dataset where a fitted model is substantially different is a shared idea between Regression Clustering and EMM. However, in Regression Clustering the subsets are not necessarily coherent, easy to describe subgroups: the goal is not to explore exceptionalities, but to give a well-fitting partition.

A similar caveat holds for the well-known Classification And Regression Trees (Breiman et al. 1984), where a nominal or numerical target concept is assigned a different class or outcome depending on conditions on attributes. While the recursive partitioning given by the tree ensures that every path from the root to a leaf constitutes a coherent, easy to describe subgroup, there is again no explicit search for exceptional-

ities. A partition that performs well is enough, and if multiple exceptional phenomena that happen to have similar effects on the target are lumped together in one cell of the partition, the CART algorithm is happy while the Exceptional Model Miner is not.

Contrary to ordinary decision trees, where the classes are found in the leaves of the tree and the internal nodes merely contain conditions for classification, a Predictive Clustering Tree (PCT) (Blockeel et al. 1998) has each internal node and each leaf corresponding to a cluster. A cluster is represented by a prototype, and a distance measure is assumed that computes the distance between prototypes hence clusters. Given all this, the decision tree algorithm is adapted to select in each node the condition maximizing the distance between the clusters in its children. Defining a quality measure that finds an optimal separation between a subset of the data and its complement, is a common concept in PCT and EMM. However, the goal of PCT is not to find global exceptionalities, but rather find a partition of the data that is optimal in some sense.

The work on PCTs has been generalized to concern the general problem of mining on a dataset with structure on the output classes, whether this structure takes the form of dependencies between classes (tree-shaped hierarchy, directed acyclic graph) or internal relations between classes (sequences). A tree ensemble method working on such data was proposed by Kocev et al. (2013). Such a method is able to give different predictions for parts of the dataset that behave differently from the norm. However, contrary to EMM, there is no explicit identification of the deviating subgroup and model.

8 Reasons for performing Exceptional Model Mining

So far in this paper we have seen what EMM is, how we can define EMM instances in a sensible way, and what kind of subgroups we can find with it. All this obviously raises the question why we would need EMM. In this section, we give one trivial and two more complicated answers to that question.

For starters, there's the trivial reason to perform EMM: we learn things about our data. Extracting pieces of information from a raw dataset is the core business of data mining, and it should not be thought of lightly if a method does merely that. As we have seen in Sect. 6, each subgroup one can find with EMM is such a coherent nugget of information. Those real-life nuggets are far more actionable for a domain expert than the raw data could ever be. Given that EMM is able to capture a richer concept of "interestingness" than conventional SD, EMM can retrieve subgroups containing more information out of the data than was possible beforehand, as long as the domain expert and the data miner together can formulate a model for the particular concept of interestingness that they strive to find.

Beyond the trivial reason, EMM is a great tool for metalearning. For example, in Sect. 5.4 we introduced an EMM instance with a classification model as target concept. Hence this instance finds subgroups for which the classification is performed in a substantially different manner than overall, which could be interesting to the researcher. Additionally, one could mine explicitly on a metadataset crafted from the results of a classification run. Suppose one is interested in predicting a numerical variable, for instance the number of days a court case will take to resolve. Having

trained and tested a classifier, we end up with a metadataset of court cases, each with the real number of days and the predicted number of days. We can now use these real and predicted numbers as the two targets in an EMM run, for instance using the correlation model from Sect. 5.1. This EMM run will result in coherent subsets of the data for which the predictions of our classifier are particularly good or bad, which is potentially very useful information for further development or finetuning of the classification algorithm.

Lastly, the subgroups found through EMM may be directly applicable in a setting that is less exploratory and more oriented towards a concrete goal. The EMM instance with a Bayesian network model as target concept, which we discussed in Sect. 5.5, is a good example. While the original goal of the EMM instance was simply to find subgroups for which the conditional dependence relations between the targets were unusual, the subgroups have been shown capable to improve multi-label SVM classifiers (Duivesteijn et al. 2012b), though it didn't work as well for decision trees. The main idea is that every subgroup can be seen as a binary attribute of the dataset, indicating whether the record is covered by the subgroup. These binary attributes highlight regions in the dataset where the labels interact in an unusual manner, so employing them in the learning phase may improve a multi-label classifier. Even though predictiveness was not considered at all when the subgroups were found, the classifier performance of SVM methods improved when these additional attributes were available. The following section details experiments providing evidence for similar global applicability of subgroups, found through EMM with the general linear regression model discussed in Sect. 5.6.

8.1 Subgroup-reinforced general linear regression modeling

In Sect. 5.6, we have discussed EMM with the General Linear Regression model class, which can be denoted by:

$$\ell_m^i = \beta_0 + \beta_1 \ell_1^i + \ldots + \beta_{m-1} \ell_{m-1}^i + \varepsilon^i \tag{5}$$

In Sect. 6.6, we have discussed subgroups found with this EMM instance on three datasets. In this section, we explore whether the quality of the global regression model can be improved by incorporating each of the found subgroups. Incorporation of a subgroup G_D is achieved by adding m new terms to the model: the subgroup indicator variable D itself, and an interaction term $D \times \ell_j$ for each explanatory variable ℓ_j in the original model. Hence, the linear regression model with an incorporated subgroup G_D can be denoted by:

$$\ell_m^i = \beta_0 + \beta_1 D^i + \beta_2 \ell_1^i + \beta_3 \ell_1^i D^i + \ldots + \beta_{2m-2} \ell_{m-1}^i + \beta_{2m-1} \ell_{m-1}^i D^i + \varepsilon^i \tag{6}$$

Here, we write D^i as shorthand for $D\left(a_1^i, \ldots, a_k^i\right)$. We deviate from the enhanced model definition on the Giffen dataset. Here we only added one interaction term,

for $D \times \%\Delta p_{i,t}$ because the subgroups were found using Cook's distance for the coefficient of $\%\Delta p_{i,t}$ only.

We would like to quantify whether the enhanced model of Eq. (6) constitutes an improvement over the original model of Eq. (5). Now, obviously, since the enhanced model encompasses all the terms of the original model (and adds some), its goodness-of-fit will invariably be better, so that comparison would be unfair. Instead, we measure model quality by the adjusted coefficient of multiple determination (adjusted R^2). Adjusted R^2 is defined as:

$$R_a^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{\text{SSE}}{\text{SST}}$$

where SSE stands for the Sum of Squared Errors and SST denotes the Sum of Squares Total. The adjusted coefficient of multiple determination may become smaller when an extra explanatory variable is added, because the decrease in SSE may be more than offset by the loss of a degree of freedom in the denominator $n - p$. We compare the adjusted R^2 of the original global model with the adjusted R^2 of the global model with the subgroup variables added. We also count how often coefficients of the subgroup variables were significant at the $\alpha = 0.05$ significance level. In addition to the datasets from Sect. 6.6, we perform experiments on three other datasets.

The *Ames Housing* dataset contains information from the Ames Assessor's Office used in computing assessed values for individual residential properties sold in Ames, Iowa from 2006 to 2010. The global model is

$$Price = -108225.05 + 1.93 \times LotArea + 44201.87 \times Quality$$

where *Price* is the sales price of the house in dollars, *Lot Area* is the lot size in square feet, and *Quality* rates the overall material and finish of the house on a scale from 1 to 10.

The *Auction* dataset was analyzed in Rezende (2008). It concerns eBay auctions of Apple iPod mini players from June 27 to July 18, 2006. The goal is to model the final price reached in the auction in terms of auction, seller, and product characteristics. The global model is

$$\begin{aligned} Price = \; & 1193.38 + 7.95 \times Nbid + 0.13 \times PositiveFeedback - 0.00 \times Time \\ & - 0.00 \times FeedbackScore - 0.10 \times Memory + 0.66 \times ResPrice \end{aligned}$$

where *Price* is the final price of the auction in US dollars, *Nbid* is the number of distinct people who bid in the auction, *PositiveFeedback* is the seller's positive feedback percentage (the coefficient is nonzero from the fourth decimal place), *Time* is the time of he final bid expressed in seconds after Dec. 31 1969, 22:00:00 PDT (the coefficient is nonzero from the fifth decimal place), *FeedbackScore* is the seller's feedback score, *Memory* is the reported memory of the iPod in gigabytes, and *ResPrice* is the auction reservation price in US dollars.

Finally, the *Wine* dataset was analyzed in Costanigro et al. (2009). It is derived from 10 years (1991–2000) of tasting ratings reported in the Wine Spectator Magazine

Table 6 Results of subgroup-reinforced general regression modeling experiments

Ω	Subgroups in			R_a^2 improvement	
	$\mathcal{G}$	$\mathcal{R}$	$\mathcal{C}$	Average	Maximum
Ames Housing	50	50	50	0.01462	0.04387
Auction	26	26	20	0.02453	0.08520
EAEF	50	50	47	0.01492	0.04993
Giffen	50	35	11	0.00149	0.02260
PC486	6	6	6	0.03989	0.08447
Wine	50	50	42	0.01210	0.05910

The set of subgroups available for testing is denoted by $\mathcal{G}$, the set of subgroups whose inclusion in the global model resulted in an increase of R_a^2 is denoted by $\mathcal{R}$, and the set of subgroups for which at least one of the coefficients (of D and $D \times \ell_i$) was significant is denoted by $\mathcal{C}$. The last two columns detail the average and maximum improvement of R_a^2 made when replacing the original model of Eq. (5) with the enhanced model of Eq. (6)

(online version) for California and Washington red wines. Our analysis uses a random sample of size 5000 from the original data. For a detailed description of the data we refer to Costanigro et al. (2009). The global model is

$$Price = -186.61 - 0.0002 \times Cases + 2.35 \times Score + 5.51 \times Age$$

where *Price* is the retail price suggested by the winery, *Score* is the score from the Wine Spectator, *Age* is the years of aging before commercialization, and *Cases* is the number of cases produced (in thousands).

The results of these model-enhancing experiments can be found in Table 6. When running the original EMM algorithm with the general linear regression model, we had restricted the number of reported subgroups to the top-50, so no more subgroups were tested here. The three leftmost columns show that, for almost all subgroups, the enhanced model has an increased R_a^2; taken together, the additional terms in the regression model improve the model more than would be expected by chance. This holds for all subgroups considered here on all datasets, except for a minority (30 %) of the subgroups considered on the *Giffen* dataset. Additionally, for each dataset except for *Giffen*, for a large majority of the subgroups, at least one of the additional terms in the enhanced regression model has *by itself* a significant ($\alpha = 0.05$) coefficient. Together, these observations provide evidence for our belief that the subgroups found through EMM are not only interesting as nuggets of information, but also potentially relevant for enhancing global modeling.

9 Conclusions

We have introduced EMM, a general framework to find subgroups of the data where something exceptional, something interesting is going on. These subgroups are not just any subset of the data: they must be coherent records in the dataset, covered

by a succinct description in terms of conditions on attributes within the dataset. The attributes that can be used for such a description are strictly separated from the target attributes, which are used to evaluate the subgroups on. Hence, EMM can be seen as an extension of SD, incorporating a more complex target.

Commonly, in traditional SD, the distribution of a single attribute is used as target concept. In EMM, the target concept is a model over several attributes. We have discussed several model classes: correlation, association, simple linear regression, classification, Bayesian networks, and general linear regression models. For each such model class we have developed a quality measure: a function that extracts relevant model characteristics, and from those characteristics computes a number quantifying how exceptional a subgroup is. A subgroup is considered exceptional when the model learned from the data belonging to the subgroup differs substantially, either from the model learned from the data belonging to its complement, or from the model learned from the overall dataset (for more on this choice, see Sect. 3.2.2). An EMM run results in succinct descriptions of subgroups, where for instance two targets are unusually correlated, or where a classifier performs exceptionally good or bad, or where the conditional dependence relations between several targets deviate from the norm.

We have discussed experimental results for each of the introduced model classes. Among the most striking results are the coherent regions within Europe found on the *Mammals* data (see Sect. 6.5) with the Bayesian Network model, where animals depend on each other in a substantially different way, and the strong real-life evidence for the Giffen effect (see Sect. 6.6.1) found with the General Linear Regression model, where poor households in the Chinese province Hunan displayed a positive price elasticity of demand for rice. Apart from merely finding such exceptional subgroups, we have argued that EMM is an excellent tool for metalearning, and found subgroups can be employed to enhance global modeling: their incorporation can improve the performance of a multi-label classifier performance, and the goodness-of-fit of a regression model.

EMM is in many respects a white box system. When employing an EMM instance on a particular domain, it is fairly simple to convey to a domain expert what kind of exceptionality is being sought after (by means of agreeing on the model class). The resulting subgroups are conjunctions of a few conditions on single attributes, which should be simple to interpret for the expert. Depending on the model class, a domain expert may also be able to properly investigate the discrepancies in fitted models; for instance in the case of a correlation or regression model this may enrich the expert's understanding of the result, but in the case of a Bayesian network fitted on a hundred animals it probably will not. We expect that deploying existing EMM instances in, or developing new EMM instances for, other fields, could lead to many fruitful collaborations between data miners and experts in those fields.

Acknowledgments This research is supported in part by the Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876 "Providing Information by Resource-Constrained Analysis", project C1, and in part by the Netherlands Organisation for Scientific Research (NWO) under project number 612.065.822 (Exceptional Model Mining).

References

Agresti A (1990) Categorical data analysis. Wiley, New York

Aidt T, Tzannatos Z (2002) Unions and collective bargaining. The World Bank, Washington, DC

Anglin PM, Gençay R (1996) Semiparametric estimation of a hedonic price function. J Appl Econ 11(6):633–648

Atzmüller M, Lemmerich F (2009) Fast subgroup discovery for continuous target concepts. In: Proceedings of ISMIS, pp 35–44

Bay SD, Pazzani MJ (2001) Detecting group differences: mining contrast sets. Data Min Knowl Discov 5(3):213–246

Blockeel H, De Raedt L, Ramon J (1998) Top-down induction of clustering trees. In: Procedings of ICML, pp 55–63

Boley M, Grosskreutz H (2009) Non-redundant subgroup discovery using a closure system. In: Proceedings of ECML/PKDD, vol 1, pp 179–194

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey

de Campos LM, Fernández-Luna JM, Huete JF (2004) Bayesian networks and information retrieval: an introduction to the special issue. Inf Process Manag 40(5):727–733

Carmona CJ, González P, del Jesus MJ, Herrera F (2010) NMEEF-SD: non-dominated multiobjective evolutionary algorithm for extracting fuzzy rules in subgroup discovery. IEEE Trans Fuzzy Syst 18(5):958–970

Chao C, Velicer C, Slezak JM, Jacobsen SJ (2009) Correlates for completion of 3-dose regimen of HPV vaccine in female members of a managed care organization. Mayo Clin Proc 84(10):864–870

Cook RD (1977) Detection of influential observation in linear regression. Technometrics 19(1):15–18

Cook RD, Weisberg S (1980) Characterizations of an empirical influence function for detecting influential cases in regression. Technometrics 22(4):495–508

Cook RD, Weisberg S (1982) Residuals and influence in regression. Chapman & Hall, London

Costanigro M, Mittelhammer RC, McCluskey JJ (2009) Estimating class-specific parametric models under class uncertainty: local polynomial regression clustering in an hedonic analysis of wine markets. J Appl Econ 24:1117–1135

Davis GA (2003) Bayesian reconstruction of traffic accidents. Law Probab Risk 2:69–89

Díez FJ, Mira J, Iturralde E, Zubillaga S (1997) DIAVAL, a Bayesian expert system for echocardiography. Artif Intell Med 10:59–73

Dong G, Li J (1999) Efficient mining of emerging patterns: discovering trends and differences. In: Proceedings of KDD, pp 43–52

Dougherty C (2011) Introduction to econometrics, 4th edn. Oxford University Press, Oxford

Duivesteijn W, Feelders A, Knobbe AJ (2012) Different slopes for different folks—mining for exceptional regression models with Cook's distance. In: Proceedings of KDD, pp 868–876

Duivesteijn W, Knobbe AJ, Feelders A, van Leeuwen M (2010) Subgroup discovery meets Bayesian networks—an exceptional model mining approach. In: Proceedings of ICDM, pp 158–167

Duivesteijn W, Loza Mencía E, Fürnkranz J, Knobbe AJ (2012) Multi-label LeGo—enhancing multi-label classifiers with local patterns. In: Proceedings of IDA, pp 114–125

Friedman J, Fisher N (1999) Bump-hunting in high-dimensional data. Stat Comput 9(2):123–143

Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian networks to analyze expression data. J Comput Biol 7(3/4):601–620

Galbrun E, Miettinen P (2012) From black and white to full color: extending redescription mining outside the Boolean world. Stat Anal Data Min 5(4):284–303

Garriga GC, Heikinheimo H, Seppänen JK (2007) Cross-mining binary and numerical attributes. In: Proceedings of ICDM, pp 481–486

Gallo A, Miettinen P, Mannila H (2008) Finding subgroups having several descriptions: algorithms for redescription mining. In: Proceedings of SDM, pp 334–345

Gentleman JF, Wilk MB (1975) Detecting outliers II: supplementing the direct analysis of residuals. Biometrics 31:387–410

Goodman LA (1970) The multivariate analysis of qualitative data: interaction among multiple classifications. J Am Stat Assoc 65:226–256

Grosskreutz H, Rüping S (2009) On subgroup discovery in numerical domains. Data Min Knowl Discov 19(2):210–226

Hand DJ, Adams NM, Bolton RJ (2002) Pattern detection and discovery, vol 2447. Lecture notes in computer science, Springer, Berlin
Heckerman D, Geiger D, Chickering DM (1995) Learning Bayesian networks: the combination of knowledge and statistical data. Mach Learn 20:197–243
Heikinheimo H, Fortelius M, Eronen J, Mannila H (2007) Biogeography of European land mammals shows environmentally distinct and spatially coherent clusters. J Biogeogr 34(6):1053–1064
Herrera F, Carmona CJ, González P, del Jesus MJ (2011) An overview on subgroup discovery: foundations and applications. Knowl Inf Syst 29(3):495–525
Hochberg Y, Tamhane A (1987) Multiple comparison procedures. Wiley, New York
Jensen RT, Miller NH (2008) Giffen behavior and subsistence consumption. Am Econ Rev 98(4):1553–1577
del Jesús MJ, González P, Herrera F, Mesonero M (2007) Evolutionary fuzzy rule induction process for subgroup discovery: a case study in marketing. IEEE Trans Fuzzy Syst 15(4):578–592
Jorge AM, Azevedo PJ, Pereira F (2006) Distribution rules with numeric attributes of interest. In: Proceedings of PKDD, pp 247–258
Klösgen W (1996) Explora: a multipattern and multistrategy discovery assistant. In: Advances in knowledge discovery and data mining. pp 249–271
Klösgen W (1998) Deviation and association patterns for subgroup mining in temporal, spatial, and textual data bases. In: Rough sets and current trends in computing. Springer, pp 1–18
Klösgen W (1999) Applications and research problems of subgroup mining. In: Proceedings of ISMIS, pp 1–15
Klösgen W (2002) Subgroup discovery. In: Handbook of data mining and knowledge discovery, chap. 16.3. Oxford University Press, New York
Knobbe AJ, Feelders A, Leman D (2012) Exceptional model mining. In: Data mining: foundations and intelligent paradigms, intelligent systems reference library, vol 24, pp 183–198
Knuth DE (1998) The art of computer programming, vol. 3: sorting and searching, 2nd edn. Addison-Wesley, Reading
Kocev D, Vens C, Struyf J, Džeroski S (2013) Tree ensembles for predicting structured outputs. Pattern Recogn 46(3):817–833
Kohavi R (1995) The power of decision tables. In: Proceedings of ECML, pp 174–189
van de Koppel E, Slavkov I, Astrahantseff K, Schramm A, Schulte J, Vandesompele J, de Jong E, Dzeroski S, Knobbe AJ (2007) Knowledge discovery in neuroblastoma-related biological data. In: Data mining in functional genomics and proteomics workshop at PKDD 2007, Warsaw, Poland, pp 45–56
Kralj Novak P, Lavrač N, Webb GI (2009) Supervised descriptive rule discovery: a unifying survey of contrast set, emerging pattern and subgroup mining. J Mach Learn Res 10:377–403
Kriegel H-P, Kröger P, Schubert E, Zimek A (2012) Outlier detection in arbitrarily oriented subspaces. In: Proceedings of ICDM, pp 379–388
Lavrač N, Flach P, Zupan B (1999) Rule evaluation measures: a unifying view. In: Proceedings of the ninth international workshop on inductive logic programming. Lecture notes in artificial intelligence, vol 1634, pp 174–185
Lavrač N, Kavšek B, Flach PA, Todorovski L (2004) Subgroup discovery with CN2-SD. J Mach Learn Res 5:153–188
van Leeuwen M (2010) Maximal exceptions with minimal descriptions. Data Min Knowl Discov 21(2):259–276
van Leeuwen M, Knobbe AJ (2011) Non-redundant subgroup discovery in large and complex data. In: Proceedings of ECML/PKDD, vol 3, pp 459–474
van Leeuwen M, Knobbe AJ (2012) Diverse subgroup set discovery. Data Min Knowl Discov 25(2):208–242
Leman D, Feelders A, Knobbe AJ (2008) Exceptional model mining. In: Proceedings of ECML/PKDD, vol 2, pp 1–16
Lemmerich F, Becker M, Atzmüller M (2012) Generic pattern trees for exhaustive exceptional model mining. In: Proceedings of ECML/PKDD, vol 2, pp 277–292
Mampaey M, Nijssen S, Feelders A, Knobbe AJ (2012) Efficient algorithms for finding richer subgroup descriptions in numeric and nominal data. In: Proceedings of ICDM, pp 499–508
Marshall A (1895) Principles of economics. MacMillan and co, New York
Meeng M, Knobbe AJ (2011) Flexible enrichment with Cortana—Software Demo. In: Proceedings of Benelearn, pp 117–119
Mitchell-Jones T et al (1999) The atlas of European mammals. Poyser natural history. Poyser, London

Moore D, McCabe G (1993) Introduction to the practice of statistics. WH Freeman and Company, New York
Morik K, Boulicaut JF, Siebes A (2005) Local pattern detection. Lecture notes in computer science, vol 3539, Springer, Heidelberg
Neil M, Fenton N, Tailor M (2005) Using Bayesian networks to model expected and unexpected operational losses. Risk Anal 25(4):963–972
Neter J, Kutner M, Nachtsheim CJ, Wasserman W (1966) Applied linear statistical models. WCB McGraw-Hill, Boston
Paine RT (1966) Food web complexity and species diversity. Am Nat 100(910):65–75
Ramakrishnan N, Kumar D, Mishra B, Potts M, Helm RF (1995) Turning CARTwheels: an alternating algorithm for mining redescriptions. In: Proceedings of KDD, pp 837–844
Rezende L (2008) Econometrics of auctions by least squares. J Appl Econ 23:925–948
Scholz M (2005) Knowledge-based sampling for subgroup discovery. In: Morik K, Boulicaut JF, Siebes A (eds) Local pattern detection. Lecture notes in computer science, vol 3539, Springer, Heidelberg, pp 171–189
Schubert E, Wolfe J, Tarnopolsky A (2004) Spectral centroid and timbre in complex, multiple instrumental textures. In: Proceedings of 8th international conference on music perception & cognition, pp 654–657
Siebes A (1995) Data surveying: foundations of an inductive query language. In: Proceedings of KDD, pp 269–274
Stengos T, Zacharias E (2006) Intertemporal pricing and price discrimination: a semiparametric hedonic analysis of the personal computer market. J Appl Econ 21:371–386
Trohidis K, Tsoumakas G, Kalliris G, Vlahavas IP (2008) Multi-label classification of music into emotions. In: Proceedings of 9th international conference on music information retrieval, pp 325–330
Umek L, Zupan B (2011) Subgroup discovery in data sets with multi-dimensional responses. Intell Data Anal 15(4):533–549
Verma T, Pearl J (1990) Equivalence and synthesis of causal models. In: Proceedings of UAI, pp 255–270
Whittaker J (1990) Graphical models in applied multivariate statistics. Wiley, New York
Wrobel S (1997) An algorithm for multi-relational discovery of subgroups. In: Proceedings of PKDD, pp 78–87
Yang G, Le Cam L (2000) Asymptotics in statistics: some basic concepts. Springer, Berlin
Zhang B (2003) Regression clustering. In: Proceedings of ICDM, pp 451–458
Zimmermann A, De Raedt L (2009) Cluster-grouping: from subgroup discovery to clustering. Mach Learn 77(1):125–159