Exceptional Model Mining

Arno Knobbe[1], Ad Feelders[2], and Dennis Leman[2]

[1] LIACS, Leiden University, Niels Bohrweg 1,
NL-2333 CA, Leiden, the Netherlands
knobbe@liacs.nl
[2] Utrecht University, P.O. box 80 089,
NL-3508 TB Utrecht, the Netherlands
{ad,dlleman}@cs.uu.nl

Abstract. In most databases, it is possible to identify small partitions of the data where the observed distribution is notably different from that of the database as a whole. In classical subgroup discovery, one considers the distribution of a single nominal attribute, and exceptional subgroups show a surprising increase in the occurrence of one of its values. In this paper, we describe *Exceptional Model Mining* (EMM), a framework that allows for more complicated target concepts. Rather than finding subgroups based on the distribution of a single target attribute, EMM finds subgroups where a model fitted to that subgroup is somehow exceptional. We discuss regression as well as classification models, and define quality measures that determine how exceptional a given model on a subgroup is. Our framework is general enough to be applied to many types of models, even from other paradigms such as association analysis and graphical modeling.

1 Introduction

By and large, subgroup discovery has been concerned with finding regions in the input space where the distribution of a single target variable is substantially different from its distribution in the whole database [3, 4]. We propose to extend this idea to targets that are models of some sort, rather than just single variables. Hence, in a very general sense, we want to discover subgroups where a model fitted to the subgroup is substantially different from that same model fitted to the entire database [5].

As an illustrative example, consider the simple linear regression model

$$P_i = a + bS_i + e_i$$

where P is the sales price of a house, S the lot size (measured, say, in square meters), and e the random error term (see Fig. 1 and Section 4 for an actual dataset containing such data). If we think the location of the house might make a difference for the price per square meter, we could consider fitting the same model to the subgroup of houses on a desirable location:

$$P_i = a_D + b_D S_i + e_i,$$

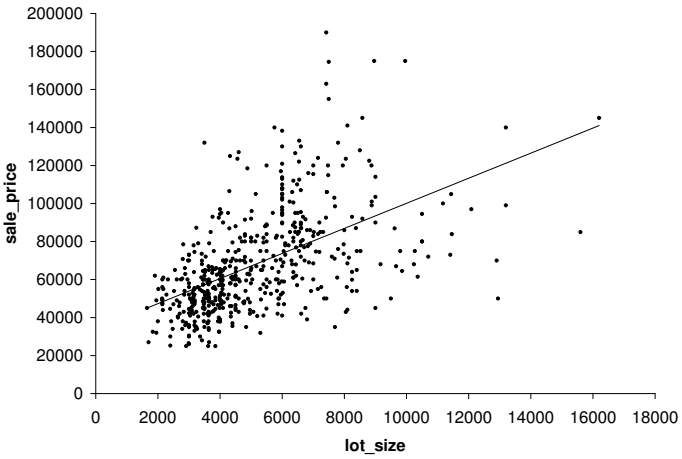

Fig. 1. Scatter plot of *lot_size* and *sales_price* for the housing data.

where the subscript D indicates we are only considering houses on a desirable location. To test whether the slope for desirable locations is significantly different, we could perform a statistical test of $H_0 : b = b_D$, or more conveniently, $H_0 : b_D = b_{\bar{D}}$, where $\bar{D}$ denotes the complement of D.

In the above example, we came up ourselves with the idea that houses on a desirable location might have a different slope in the regression model. The main idea presented in this paper is that we can find such groups automatically by using the subgroup discovery framework. Hence, the subgroups are not limited to simple conditions based on a single variable. Their description may involve conjunctions of conditions, and in case of multi-relational data, existential quantification and aggregation as well. In the general case of simple linear regression, we could be looking for subgroups G where the slope b_G in

$$y_i = a_G + b_G x_i + e_i,$$

is substantially different from the slope $b_{\bar{G}}$. The search process only involves the subgroups; the variables y and x are assumed to be determined by the question of the user, that is, they are fixed.

We have stated that the objective is to find subgroups where a model fitted to the subgroup is substantially different from that same model fitted to the entire database. This statement is deliberately general: we can use different types of models in this scheme, and for each type of model we can consider several measures of difference. In this paper we describe a number of model classes and quality measures that can be useful. All these methods have been implemented in the Multi-Relational Data Mining system Safarii [6].

This paper is organized as follows. In Section 2, we introduce some notation that is used throughout the paper, and define the subgroup discovery and exceptional model mining framework. In Section 3, we give examples of three basic

types of models for exceptional model mining: correlation, regression and classification. We also propose appropriate quality measures for the types of models discussed. In Section 4, we present the results of exceptional model mining applied to two real-life datasets. Finally, we draw conclusions in Section 5.

2 Exceptional Model Mining

We assume that the database d is a bag of labelled objects $i \in D$, referred to as *individuals*, taken from a domain D. We refer to the size of the database as $N = |d|$. At this point, we do not fix the nature of individuals, be it propositional, relational, or graphical, etc. However, each description of an individual includes a number of attributes $x_1, ..., x_k$ and optionally an output attribute y. These attributes are used in fitting models to subgroups of the data. In regular subgroup discovery, only the y attribute is used, which is typically binary.

We make no assumptions about the syntax of the pattern language, and treat a pattern simply as a function $p : D \to \{0, 1\}$. We will say that a pattern p *covers* an individual i *iff* $p(i) = 1$.

Definition 1 (Subgroup). *A* subgroup *corresponding to a pattern p is the set of individuals $G_p \subseteq d$ that are covered by p: $G_p = \{i \in d | p(i) = 1\}$.*

Definition 2 (Complement). *The* complement *of a subgroup G_p is the set of individuals $\bar{G}_p \subseteq d$ that are* not *covered by p: $\bar{G}_p = d \backslash G_p$.*

When clear from the context, we will omit the p from now on, and simply refer to a subgroup and its complement as G and $\bar{G}$. We use n and $\bar{n}$ to denote the size of G and $\bar{G}$, respectively. In order to judge the quality of candidate patterns in a given database, a *quality measure* needs to be defined. This measure determines for each pattern in a pattern language $\mathcal{P}$ how interesting (exceptional) a model induced on the associated subgroup is.

Definition 3 (Quality Measure). *A* quality measure *for a pattern p is a function $\varphi_d : \mathcal{P} \to \mathbb{R}$ that computes a unique numeric value for a pattern p, given a database d.*

Subgroup discovery [3] is a data mining framework aimed at discovering patterns that satisfy a number of user-specified inductive constraints. These constraints typically include an interestingness constraint $\varphi(p) \geq t$, as well as a minimum support threshold $n \geq minsup$ that guarantees the relative frequency of the subgroups in the database. Further constraints may involve properties such as the complexity of the pattern p. In most cases, a subgroup discovery algorithm will traverse a search lattice of candidate patterns in a top-down, general-to-specific fashion. The structure of the lattice is determined by a *refinement operator* $\rho : \mathcal{P} \to 2^{\mathcal{P}}$, a syntactic operation which determines how simple patterns can be extended into more complex ones by atomic additions. In our application (and most others), the refinement operator is assumed to be a *specialisation operator*: $\forall q \in \rho(p) : p \succeq q$ (p is more general than q).

The actual search strategy used to consider candidates is a parameter of the algorithm. We have chosen the *beam search* strategy [14], because it nicely balances the benefits of a greedy method with the implicit parallel search resulting from the beam. Beam search effectively performs a level-wise search that is guided by the quality measure φ. On each level, the best-ranking w patterns are refined to form the candidates for the next level. This means that although the search will be targeted, it is less likely to get stuck in a local optimum, because at each level alternatives are being considered. The search is further bounded by complexity constraints and the *minsup* constraint. The end-result is a ranked list of patterns (subgroups) that satisfy the inductive constraints.

In the case of regular subgroup discovery, with only a single discrete target variable, the quality measure of choice is typically a measure for how different the distribution over the target variable is, compared to that of the whole database (or in fact to that of the complement). As such an unusual distribution is easily produced in small fractions of the database, the deviation is often weighed with the size of the subgroup: a pattern is interesting if it is both exceptional and frequent. Well-known examples of quality measures for binary targets are frequency, confidence, χ^2, and novelty.

The subject of this paper, exceptional model mining (EMM) [5], can now be viewed as an extension of the subgroup discovery framework. The essential difference with standard subgroup discovery is the use of more complex target concepts than the regular single attribute. Our targets are models of some sort, and within each subgroup considered, a model is induced on the attributes $x_1, ..., x_k$, and optionally y. We will define quality measures that capture how exceptional the model within the subgroup is in relation to the model induced on its complement. In the next section, we present a number of model types, and propose one or more quality measures for each. When only the subgroup itself is considered, the quality measures tend to focus on the accuracy of the model, such as the fit of a regression line, or the predictive accuracy of a classifier. If the quality measure captures the difference between the subgroup and its complement, it is typically based on a comparison between more structural properties of the two models, such as the slope of the regression lines, or the make-up of the classifiers (e.g. size, attributes used).

Example 1. Consider again the housing dataset (Fig. 1). Individuals (houses) are described by a number of attributes such as the number of bathrooms or whether the house is located at a desirable location. An example of a pattern (and associated subgroup G) would be:

$$p : nbath \geq 2 \wedge drive = 1$$

which covers 128 houses (about 23% of the data). Its complement (which is often only considered implicitly) is

$$\bar{p} : \neg nbath \geq 2 \vee \neg drive = 1$$

The typical refinement operator will add a single condition on any of the available attributes to the conjunction. In this example, target models are defined over the

two attributes $x = lot\_size$ and $y = sales\_price$. Note that these two attributes are therefore not allowed to appear in the subgroup definitions. One possibility is to perform the linear regression of y on x. As a quality measure φ_d, we could consider the absolute difference in slope between the two regression lines fitted to G and $\bar{G}$. In Section 3.2, we propose a more sophisticated quality measure for the difference in slope, that implicitly takes into account the supports n and $\bar{n}$, and thus the significance of the finding.

3 Model Classes

In this section, we discuss simple examples of three classes of models, and suggest quality measures for them. As an example of a model without an output attribute, we consider the correlation between two numeric variables. We discuss linear regression for models with a numeric output attribute, and two simple classifiers for models with discrete output attributes.

3.1 Correlation models

As an example of a model without an output attribute, we consider two numeric variables x_1 and x_2, and their linear association as measured by the correlation coefficient ρ. We estimate ρ by the sample correlation coefficient r:

$$r = \frac{\sum(x_1^i - \bar{x}_1)(x_2^i - \bar{x}_2)}{\sqrt{\sum(x_1^i - \bar{x}_1)^2 \sum(x_2^i - \bar{x}_2)^2}}$$

where x^i denotes the i^{th} observation on x, and $\bar{x}$ denotes its mean.

Absolute difference between correlations (φ_{abs}). A logical quality measure is to take the absolute difference of the correlation in the subgroup G and its complement $\bar{G}$, that is

$$\varphi_{abs}(p) = |r_G - r_{\bar{G}}|$$

The disadvantage of this measure is that it does not take into account the size of the groups, and hence does not do anything to prevent overfitting. Intuitively, subgroups with higher support should be preferred.

Entropy (φ_{ent}). As an improvement of φ_{abs}, the following quality function weighs the absolute difference between the correlations with the *entropy* of the split between the subgroup and its complement. The entropy captures the information content of such a split, and favours 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 $H(p)$ is defined (in this context) as:

$$H(p) = -n/N \lg n/N - \bar{n}/N \lg \bar{n}/N$$

The quality measure φ_{ent} is now defined as:

$$\varphi_{ent}(p) = H(p) \cdot |r_G - r_{\bar{G}}|$$

Significance of correlation difference (φ_{scd}). A more statistically oriented approach to prevent overfitting is to perform a hypothesis test on the difference between the correlation in the subgroup and its complement. Let ρ_p and $\rho_{\bar{p}}$ denote the population coefficients of correlation for p and $\bar{p}$, respectively, and let r_G and $r_{\bar{G}}$ denote their sample estimates. The test to be considered is

$$H_0 : \rho_p = \rho_{\bar{p}} \qquad \text{against} \qquad H_a : \rho_p \neq \rho_{\bar{p}}$$

We would like to use the observed significance (p-value) of this test as a quality measure, but the problem is that the sampling distribution of the sample correlation coefficient is not known in general. If x_1 and x_2 follow a bivariate normal distribution, then application of the Fisher z transformation

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

makes the sampling distribution of z' approximately normal [12]. Its standard error is given by

$$\frac{1}{\sqrt{m-3}}$$

where m is the size of the sample. As a consequence

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

approximately follows a standard normal distribution under H_0. Here z' and $\bar{z}'$ are the z-scores obtained through the Fisher z transformation for G and $\bar{G}$, respectively. If both n and $\bar{n}$ are greater than 25, then the normal approximation is quite accurate, and can safely be used to compute the p-values. Because we have to introduce the normality assumption to be able to compute the p-values, they should be viewed as a heuristic measure. Transformation of the original data (for example, taking their logarithm) may make the normality assumption more reasonable. As a quality measure we take 1 minus the computed p-value so that $\varphi_{scd} \in [0, 1]$, and higher values indicate a more interesting subgroup.

3.2 Regression Model

In this section, we discuss some possibilities of EMM with regression models. For ease of exposition, we only consider the linear regression model

$$y_i = a + bx_i + e_i, \tag{1}$$

but this is in no way essential to the methods we discuss.

Significance of Slope Difference (φ_{ssd}). Consider model (1) fitted to a subgroup G and its complement $\bar{G}$. Of course, there is a choice of distance measures between the fitted models. We propose to look at the difference in the slope b 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 a, 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 b_p be the slope for the regression function of p and $b_{\bar{p}}$ the slope for the regression function of $\bar{p}$. The hypothesis to be tested is

$$H_0 : b_p = b_{\bar{p}} \qquad \text{against} \qquad H_a : b_p \neq b_{\bar{p}}$$

We use the least squares estimate

$$\hat{b} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

for the slope b. An unbiased estimator for the variance of $\hat{b}$ is given by

$$s^2 = \frac{\sum \hat{e}_i^2}{(m-2)\sum (x_i - \bar{x})^2}$$

where $\hat{e}_i$ is the regression residual for individual i, and m is the sample size. Finally, we define our test statistic

$$t' = \frac{\hat{b}_G - \hat{b}_{\bar{G}}}{\sqrt{s_G^2 + s_{\bar{G}}^2}}$$

Although t' does not have a t distribution, its distribution can be approximated quite well by one, with degrees of freedom given by (cf. [11]):

$$df = \frac{\left(s_G^2 + s_{\bar{G}}^2\right)^2}{\frac{s_G^4}{n-2} + \frac{s_{\bar{G}}^4}{\bar{n}-2}} \tag{2}$$

Our quality measure $\varphi_{ssd} \in [0, 1]$ is once again defined as one minus the p-value computed on the basis of a t distribution with degrees of freedom given in (2). If $n + \bar{n} \geq 40$ the t-statistic is quite accurate, so we should be confident to use it unless we are analysing a very small dataset.

3.3 Classification Models

In the case of classification, we are dealing with models for which the output attribute y is discrete. In general, the attributes $x_1, ..., x_k$ 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 two relatively simple classifiers here, for reasons of simplicity and efficiency.

Logistic Regression. Analogous to the linear regression case, we consider the logistic regression model

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

where $y \in \{0, 1\}$ is a binary class label. The coefficient b 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 b indicates that an increase in x leads to an increase of $P(y = 1|x)$ and vice versa. 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_p, we fit the model

$$\text{logit}(P(y_i = 1|x_i)) = a + b \cdot p(i) + c \cdot x_i + d \cdot (p(i) \cdot x_i). \tag{3}$$

Note that

$$\text{logit}(P(y_i = 1|x_i)) = \begin{cases} (a + b) + (c + d) \cdot x_i & \text{if } p(i) = 1 \\ a + c \cdot x_i & \text{if } p(i) = 0 \end{cases}$$

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 $d = 0$ against a two-sided alternative in the model of equation (3). This is a standard test in the literature on logistic regression [12]. We refer to this quality measure as φ_{sed}.

DTM classifier. The second classifier considered is the *Decision Table Majority* (DTM) classifier [8, 7], also known as a *simple decision table*. The idea behind this classifier is to compute the relative frequencies of the y values for each possible combination of values for $x_1, \ldots, x_k$. For combinations that do not appear in the dataset, the relative frequency estimates are based on that of the whole dataset. The predicted y value for a new individual is simply the one with the highest probability estimate for the given combination of input values.

Example 2. 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 *BDeu* (Bayesian Dirichlet equivalent uniform) score, which is a measure for the performance of the DTM classifier on G, and the *Hellinger distance*, which assigns a value to the distance between the conditional probabilities estimated on G and $\bar{G}$.

BDeu score (φ_{BDeu}). The BDeu score φ_{BDeu} is a measure from Bayesian theory [2] and is used to estimate the performance of a classifier on a subgroup, 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 subgroup is.

The BDeu score is defined as

$$\prod_{x_1,\ldots,x_k} \frac{\Gamma(\alpha/q)}{\Gamma(\alpha/q + n(x_1,...,x_k))} \prod_{y} \frac{\Gamma(\alpha/qr + n(x_1,..,x_k,y))}{\Gamma(\alpha/qr)}$$

where Γ denotes the gamma function, q denotes the number of value combinations of the input variables, r the number of values of the output variable, and $n(x_1,...,x_k,y)$ denotes the number of cases with that value combination. The parameter α denotes the *equivalent sample size*. Its value can be chosen by the user.

Hellinger (φ_{Hel}). Another possibility is to use the Hellinger distance [13]. It defines the distance between two probability distributions $P(z)$ and $Q(z)$ as follows:

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

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

$$P(y \mid x_1,...,x_k)$$

for each possible value combination $x_1,...,x_k$. The overall distance measure becomes

$$\varphi_{Hel}(p) = D(\hat{P}_G, \hat{P}_{\bar{G}}) = \sum_{x_1,\ldots,x_k} \sum_{y} \left(\sqrt{\hat{P}_G(y|x_1,...,x_k)} - \sqrt{\hat{P}_{\bar{G}}(y|x_1,...,x_k)}\right)^2$$

where $\hat{P}_G$ denotes the probability estimates on G. Intuitively, we measure the distance between the conditional distribution of y in G and $\bar{G}$ for each possible combination of input values, and add these distances to obtain an overall distance. Clearly, this measure is aimed at producing subgroups for which the conditional distribution of y is substantially different from its conditional distribution in the overall database.

4 Experiments

This section illustrates exceptional model mining on two real-life datasets, using different quality measures. Although our implementation in Safarii essentially is multi-relational [6], the two dataset we present are propositional. For each test, Safarii returns a configurable number of subgroups ranked according to the quality measure of choice. The following experiments only present the best ranking subgroup and take a closer look at the interpretation of the results.

4.1 Analysis of Housing Data

First, we analyse the Windsor housing data[3] [9]. 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, *lot_size* and *sales_price*, as plotted in Fig. 1. 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:

$$\hat{y} = 34136 + 6.60 \cdot x$$

As this function shows, on average one extra square meter corresponds to a 6.6 dollar higher sales price. Given this function, one might wonder whether it is possible to find specific subgroups in the data where the price of an additional square meter is significantly less, perhaps even zero. In the next paragraphs, we show how EMM may be used to answer this question.

Significance of Correlation Difference. Looking at the restrictions defined in Section 3.1 we see that the support has to be over 25 in order to be confident about the test results for this measure. This number was used as minimum support threshold for a run of Safarii using φ_{scd}. The following subgroup (and its complement) was found to show the most significant difference in correlation: $\varphi_{scd}(p_1) = 0.9993$.

$$p_1 : drive = 1 \wedge rec\_room = 1 \wedge nbath \geq 2.0$$

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. 2. The subgroup shows a correlation of $r_G = -0.090$ compared to $r_{\bar{G}} = 0.549$ 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

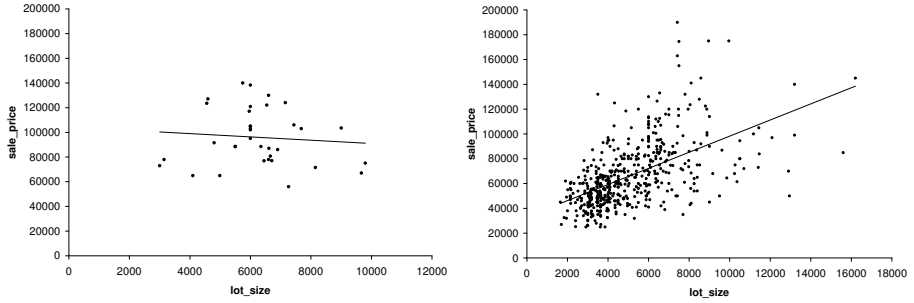

Fig. 2. Housing - φ_{scd}: Scatter plot of *lot_size* and *sales_price* for $drive = 1 \wedge rec\_room = 1 \wedge nbath \geq 2$ (left) and its complement (right).

is not a factor in the pricing. Figure 2 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

$$H_0 : b_{p_1} = 0 \qquad \text{against} \qquad H_a : b_{p_1} \neq 0,$$

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 different 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 $r = 0.129$. When we add the condition that the house should also have a driveway we arrive at the final result with $r = -0.090$. Note that adding this 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).

Significance of Slope Difference. In this section, we perform EMM on the housing data using the Significance of Slope Difference (φ_{ssd}) as the quality measure. The highest ranking subgroup consists of the 226 houses that have a

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

Table 1. Different subgroups of the housing data, and their sample correlation coefficients and supports.

Subgroup	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

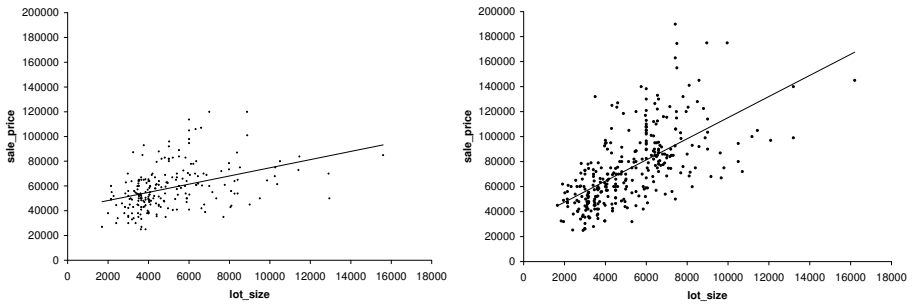

Fig. 3. Housing - φ_{ssd}: Scatter plot of $drive = 1 \wedge basement = 0 \wedge nbath \leq 1$ (left), and its complement (right).

driveway, no basement and at most one bathroom:

$$p_2 : drive = 1 \wedge basement = 0 \wedge nbath \leq 1$$

The subgroup G and its complement $\bar{G}$ (320 houses) lead to the following two fitted regression functions, respectively:

$$\hat{y} = 41568 + 3.31 \cdot x$$
$$\hat{y} = 30723 + 8.45 \cdot x$$

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

$$H_0 : b_{p_2} = b_{\bar{p}_2} \qquad \text{against} \qquad H_a : b_{p_2} \neq b_{\bar{p}_2}$$

is virtually zero. There are subgroups with a larger difference in slope, but the reported subgroup scores higher because it is quite big. Figure 3 shows the scatter plots of *lot_size* and *sales_price* for the subgroup and its complement.

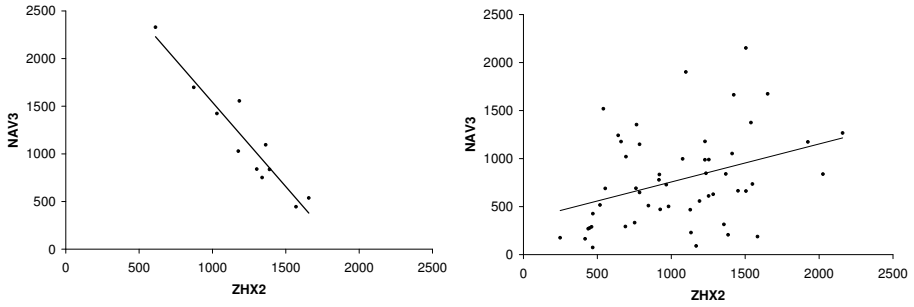

Fig. 4. Gene Expression - φ_{abs}: Scatter plot of 11_band = 'no deletion' $\wedge$ $survivaltime \leq 1919 \wedge XP\_498569.1 \leq 57$ (left; $r = -0.950$) and its complement (right; $r = 0.363$).

4.2 Analysis of Gene Expression Data

The following experiments demonstrate the usefulness of exceptional model mining in the domain of bioinformatics. In genetics, genes are organised 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 [10]. Additionally, the dataset contains clinical information about the patients, including age, sex, stage of the disease, etc.

Correlation model experiment. As a demonstration of a correlation model, we analyse the correlation between ZHX3 ('Zinc fingers and homeoboxes 2') and NAV3 ('Neuron navigator 3'), in terms of the absolute difference of correlations φ_{abs}. These genes show a very slight correlation ($r = 0.218$) in the whole dataset. The remaining attributes (both gene expression and clinical information) are available for building subgroups. 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 ($\varphi_{abs}(p_3) = 1.313$) was found with the following condition:

$$p_3 : 11\_band = \text{'no deletion'} \wedge survivaltime \leq 1919 \wedge XP\_498569.1 \leq 57$$

Figure 4 shows the plot for this subgroup and its complement with the regression lines drawn in. The correlation in the subgroup is $r_G = -0.95$ and the correlation in the remaining data is $r_{\bar{G}} = 0.363$. Note that the subgroup is very "predictable": all points are quite close to the regression line, with $R^2 \approx 0.9$.

DTM experiment. For the DTM classification experiments on the Gene Expression dataset, we have selected three binary attributes. 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* (x_1) describes whether the small arm ('p') of the first chromosome has been deleted. The second attribute, MYCN (x_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 attibute for the classification model is $NBstatus$ (y), 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 data set:

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

In order to find subgroups 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 larger subgroups, we have selected a slightly higher minimum support constraint: $minsup = 16$, which corresponds to 25% of the data. The following subgroup of 17 patients was the best found ($\varphi_{Hel} = 3.803$):

$$p_4 : prognosis = \text{'unknown'}$$

	MYCN ='amplified'	MYCN = 'not amplified'
1p_band = 'deletion'	1.0 (1)	0.833 (6)
1p_band = '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 classification experiment, we are interested in "predictable" subgroups. Therefore, we run EMM with the φ_{BDeu} measure. All other settings are kept the same. The following subgroup ($n = 16$, $\varphi_{BDeu} = -1.075$) is based on the expression of the gene RIF1 ('RAP1 interacting factor homolog (yeast)')

$$p_5 : \text{RIF1} >= 160.45$$

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

In this subgroup, 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.

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

$$\text{logit}\{P(NBstatus = \text{'relapse or deceased'})\} = a + b \cdot p + c \cdot age + d \cdot (p \cdot age).$$

We find the subgroup

$$p_6 : \text{SMPD1} \geq 840 \wedge \text{HOXB6} \leq 370.75$$

with a coverage of 33, and quality $\varphi_{sed} = 0.994$. 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.

5 Conclusions and Future Research

We have introduced exceptional model mining (EMM) as an extension of the well-known subgroup discovery framework. By focusing on models instead of single target variables, many new interesting analysis possibilities are created. We have proposed a number of model classes that can be used in EMM, and defined several quality measures for them. We illustrated the use of EMM by its application to two real datasets. Like subgroup discovery, EMM is an exploratory method that requires interaction with a user that is knowledgable in the application domain. It can provide useful insights into the subject area, but does not result in ready-to-use predictive models.

We believe there are many possibilities to extend the work presented in this paper. One could look at different models, for example naive Bayes for classification problems or graphical models for modelling the probability distribution of a number of (discrete) variables. Whatever the selected class of models, the user should specify a quality measure that relates to the more fundamental questions a user may have about the data at hand. In the case of our housing example, the choice for the difference in slope is appropriate, as it captures a relevant aspect of the data, namely a significant change in price per square meter. For similar reasons, we used the difference between the coefficients of the explanatory variable (age at diagnosis) in the subgroup and its complement as a quality measure for logistic regression models.

Specifying an appropriate quality measure that is inspired by a relevant question of the user becomes less straightforward when more complex models are considered, although of course one can always focus on some particular aspect (e.g. coefficients) of the models. However, even for sophisticated models such

as support vector machines or Bayesian networks, one can think of measures that make sense, such as the linear separability or the edit distance between two networks [15], respectively.

From a computational viewpoint, it is advisable to keep the models to be fitted simple, since many subgroups have to be evaluated in the search process. For example, fitting a naive Bayes model to a large collection of subgroups can be done quite efficiently, but fitting a support vector machine could prove to be too time consuming.

References

1. Affymetrix, `http://www.affymetrix.com/index.affx`, 1992.
2. Heckerman, D., Geiger D. and Chickering, D., Learning Bayesian Networks: The combination of knowledge and statistical data, Machine Learning, Vol. 20, pp. 179-243, 1995.
3. Klösgen, W., Handbook of Data Mining and Knowledge Discovery, chapter 16.3: Subgroup Discovery, Oxford University Press, New York, 2002.
4. Friedman, J. and Fisher, N., Bump-Hunting in High-Dimensional Data, Statistics and Computing, Vol. 9, No. 2, pp. 123-143, 1999.
5. Leman, D., Feelders, A., Knobbe, A., Exceptional Model Mining, In Proceedings of ECML PKDD 2008, LNCS 5212, Antwerp, 2008.
6. Knobbe, A., Safarii multi-relational data mining environment, `http://www.kiminkii.com/safarii.html`, 2006.
7. Knobbe, A., Ho, E., Pattern Teams, in Proceedings PKDD'06, Berlin, Springer Verlag, 2006.
8. Kohavi, R., The Power of Decision Tables, in Proceedings ECML'95, London, 1995.
9. Anglin, P.M. and Gençay, R., Semiparametric Estimation of a Hedonic Price Function, Journal of Applied Econometrics, Vol. 11, No. 6, pp. 633-648, 1996.
10. Koppel, E. van de, *et al*, Knowledge Discovery in Neuroblastoma-related Biological Data, Data Mining in Functional Genomics and Proteomics workshop at PKDD 2007, Warsaw, Poland, 2007.
11. Moore, D., McCabe, G., Introduction to the Practice of Statistics, New York, 1993.
12. Neter, J., Kutner, M., Nachtsheim, C.J., Wasserman, W., Applied Linear Statistical Models, WCB McGraw-Hill, 1996.
13. Yang, G and Le Cam, L., Asymptotics in Statistics: Some Basic Concepts, Berlin, Springer Verlag, 2000.
14. Xu, Y. and Fern, A., Learning Linear Ranking Functions for Beam Search, in Proceedings ICML'07, 2007.
15. Niculescu-Mizil, A. and Caruana, R., Inductive Transfer for Bayesian Network Structure Learning, in Proceedings of the 11th International Conference on AI and Statitics (AISTATS07), 2007.