COPYRIGHT NOTICE

This is the author's version of the work. The definitive version was published in *IEEE Transactions on Knowledge and Data Engineering* (TKDE), 2012.

DOI: `http://dx.doi.org/10.1109/TKDE.2012.58`.

Discovering Characterizations of the Behavior of Anomalous Sub-populations

Fabrizio Angiulli, Fabio Fassetti and Luigi Palopoli

Abstract—We consider the problem of discovering attributes, or properties, accounting for the a-priori stated abnormality of a group of anomalous individuals (the outliers) with respect to an overall given population (the inliers). To this aim, we introduce the notion of exceptional property and define the concept of exceptionality score, which measures the significance of a property. In particular, in order to single out exceptional properties, we resort to a form of minimum distance estimation for evaluating the badness of fit of the values assumed by the outliers compared to the probability distribution associated with the values assumed by the inliers. Suitable exceptionality scores are introduced for both numeric and categorical attributes. These scores are, both from the analytical and the empirical point of view, designed to be effective for small samples, as it is the case for outliers. We present an algorithm, called EXPREX, for efficiently discovering exceptional properties. The algorithm is able to reduce the needed computational effort by not exploring many irrelevant numerical intervals and by exploiting suitable pruning rules. The experimental results confirm that our technique is able to provide knowledge characterizing outliers in a natural manner.

Index Terms—Knowledge discovery, anomaly characterization, unbalanced data, mixed-attribute data.

1 Introduction

Assume that a data population is given characterized by a certain number of attributes. Assume, moreover, that the information is provided that a (typically small) fraction of the individuals in that data population is anomalous, but no reason whatsoever is given as to why these individuals behave anomalously. An interesting and challenging learning task consists therefore in characterizing the behavior of such anomalous individuals and the work [1] precisely considers the problem of discovering attributes that account for the (a-priori stated) abnormality of *one single* individual within a given data population.

In this paper, we extend the perspective of that approach in order to be able to deal with *groups*, or *sub-populations*, of anomalous individuals. As an example, consider a rare disease and assume a population of healthy and unhealthy human individuals is given; here, it would be very useful to single out properties characterizing the unhealthy individuals.

An *exceptional property* is an attribute characterizing the abnormality of the given anomalous group (the *outliers*) with respect to the normal data population (the *inliers*). Moreover, each property can have associated a condition, also called *explanation*, whose aim is to single out a (significant) portion of the data for which the property is indeed characterizing anomalous sub-populations.

In order to single out significant properties, we resort to minimum distance estimation methods, that are statistical methods for fitting a mathematical model to data. To judge the quality of a property, we make use of *exceptionality scores*, that are functions measuring the badness of fit of the values assumed by the outliers compared to the probability distribution associated with the values assumed by the inliers.

The exceptionality scores here defined are based on a *randomization test* using the *Pearson chi-square* criterion [2], for categorical properties, and on the *Cramér-von-Mises* criterion [3], for numerical properties. These criteria evaluate the badness of fit of a probability distribution F compared to a sample set. In particular, we employ as reference distribution F the empirical distribution function associated with the population of inliers and, as the sample set, the population of outliers. We note that the proposed exceptionality scores are specifically designed for the task at hand, in which we compare a rare population with a large population of normal individuals. Also, we present an algorithm, called EXPREX, or EXceptional PRoperty EXtractor, that automatically singles out the exceptional properties and their associated explanations.

In order to make the significance of the kind of knowledge mined by the EXPREX algorithm clear, we briefly illustrate next a real life example application scenario, which we will subsequently analyze in detail in Section 5. The example refers to the analysis of a genetic dataset (which Prof. Passarino of the Department of Cell Biology of the University of Calabria provided us), exploited in [1] as well, from which to induce some characterization of the influence of the genetic profiles of individuals upon survival at old age. This dataset has been the object of a separate investigation in [4], which resulted in a well-assessed characterization of genetic traits of longevous individuals as opposed to normal ones.

The biological interest here is in finding the influence of a gene on the longevity for individuals of a population or of a portion thereof. In particular, authors of [4] stud-

F. Angiulli, F. Fassetti and L. Palopoli are with DEIS, University of Calabria, Italy, email: {f.angiulli, f.fassetti, palopoli}@deis.unical.it

ied the influence of the genetic profiles on the longevity of males and females, separately. We notice that the kind of knowledge of interest in [4] is very similar to the one we aim to learn, since it considers a portion of data composed of males (females, resp.), which corresponds to the *explanation SEX ="M"* (*SEX ="F"*, resp.), and searches for an attribute, which corresponds to an *exceptional property*, characterizing anomalous longevous individuals (the outliers) w.r.t. normal ones (the inliers).

The results presented in [4] assert that the genes *"APOE"* and *"HSP90-β"* show a significant effect on the longevity among females. Conversely, when considering male individuals, the attributes (genes) *"APOE"*, *"APOA1"* and *"SIRT3"* behave differently on the longevous individuals with respect to the normal individuals.

We experimented our technique on this dataset by considering the individuals over a hundred years of age as the outliers and the remaning ones as the inliers. Details of the experiments are reported in Section 5. Our algorithm returned as exceptional, among others, also the properties pointed out in [4]; the following table reports the values of our exceptionality score τ on the genes singled out by [4].

Explanation: $SEX =$"F"		Explanation: $SEX =$"M"	
APOE	$\tau = 0.995$	APOE	$\tau = 0.970$
HSP90-β	$\tau = 0.972$	SIRT3	$\tau = 0.995$
		APOA1	$\tau = 0.992$

We note that in all cases the value of τ is statistically significant (larger than 0.95, corresponding to significance level $\alpha = 0.05$) in agreement with the results of [4], hereby confirming that our approach is, in fact, sensible. To sharpen the significance of these results, we note that the relevance of the gene *"APOE"* for the longevity was also pointed out in [5].

The commonality of the results produced by our approach with those described in the work [4], which have been the object of an independent study, provides evidence that the concept of exceptional property we put forth in this work is indeed significant and sensible. The recalled commonality of results, though, does not correspond to any analogies in the methods adopted here and in [4]:

- First of all, in [4] the sub-populations object of analysis are manually selected by the authors, while in our approach *properties and associated explanations are automatically singled out* by means of a search algorithm, using the exceptionality score.
- Also, in order to assess the statistical significance of the induced knowledge, we employ different criteria than in [4]. This is motivated by our need to construct a generally applicable method, which behaves properly, in particular, with *very small samples* (as outlier sets usually are).

Further evidence for the significance of our approach stems from the results we obtained from the experimental campaign we carried out using our technique which are illustrated in full in Section 5.

The rest of the work is organized as follows. In Section 2 we present the exceptionality scores and the definition of exceptional property associated with a group of outliers. In Section 3 we discuss some literature related to this work and point out analogies and differences. In Section 4 we describe the EXPREX algorithm for detecting exceptional properties with associated explanations. Then, in Section 5, we present experimental results conducted by using the EXPREX algorithm. Finally, in Section 6, we draw our conclusions.

2 Exceptional Property

In this section the concept of exceptional property is introduced and the exceptional property discovery problem is stated. The section is organized as follows. We start by giving some preliminary definitions (Section 2.1), then we introduce the exceptionality score for categorical properties (Section 2.2) and for numerical properties (Section 2.3), and, finally, we formally define the *exceptional property discovery problem* (Section 2.4).

2.1 Preliminary definitions

An *attribute* a is an identifier with an associated domain, denoted as $Dom(a)$, which can be either categorical or numerical. Given a set of attributes $A = \{a_1, \ldots, a_d\}$, $Dom(A)$ denotes the domain $Dom(a_1) \times \cdots \times Dom(a_d)$.

An *object* t on a set of attributes A is a d-uple $\langle v_1, \ldots, v_d \rangle$ of $Dom(A)$ (that is, each v_i belongs to $Dom(a_i)$). In the following, $t[a_i]$ denotes the value v_i.

A *dataset* D on a set of attributes A, is a bag of objects on A.

Let A be a set of attributes. A *condition* on A is an expression of the form $a \in [l, u]$, where (i) $a \in A$, (ii) $l, u \in Dom(a)$, and (iii) $l \leq u$, if a is numeric, and $l = u$, if a is categorical. If $l = u$, the interval $I = [l, u]$ is sometimes abbreviated as u and the condition as $a \in I$ or $a = I$.

Let D be a dataset on a set of attributes A and let c be a condition $a \in [l, u]$ on A. An object o of D satisfies the condition c, if and only if $o[a]$ equals l, if a is categorical, or $l \leq o[a] \leq u$, if a is numerical. This notion can be extended to a set of conditions. In particular, o satisfies a set of conditions C if and only if o satisfies each condition $c \in C$.

Let D be a dataset on a set of attributes A, and let C be a set of conditions on A. In the following, D_C denotes the dataset consisting of the objects $t \in D$ satisfying C, and $D[A']$ denotes the projection of the dataset D on the set of attributes $A' \subseteq A$.

Definition 1 (Exceptionality score): Given two datasets D^o (the *outlier dataset*) and D^i (the *inlier dataset*) on a set of attributes A, an (*exceptionality*) *score* of D^o w.r.t. D^i on the attribute $a \in A$, denoted as $\tau_a(D^o, D^i)$ (or simply τ, if no ambiguity arises), measures the badness of fit of the values in the set $D^o[a]$ compared to the probability distribution associated with the values in the set $D^i[a]$.

The smaller the value τ, the more likely the involved values come from the same distribution. With the score τ, a threshold value θ_τ is associated (which also may depend on a given significance level α) representing the value above which we can reject the hypothesis that data from D^o comes from the same distribution as those in D^i. If $\tau_a(D^o, D^i) > \theta_\tau$ then we say that a is an *exceptional property* for D^o w.r.t. D^i according to τ. □

Intuitively, the worse the values in $D^o[a]$ fit the distribution of the values in $D^i[a]$, the better the attribute a characterizes the outlier sub-population.

In particular, in order to define proper exceptionality score functions, we resort to minimum distance estimation methods, that are statistical methods for fitting a mathematical model to data. In the following, we introduce the exceptionality score functions we adopt in this paper. Due to the different nature of their domain, we distinguish between categorical and numerical attributes.

2.2 Categorical properties

Consider a categorical attribute. A well-known statistical test used for minimum distance estimation on categorical domains is the *Pearson chi-square test* χ^2 [2], a common application of which is realized by taking the outcomes of a categorical variable as events.

In order to build a proper exceptionality score, we use the goodness-of-fit test, which establishes if the observed frequency distribution differs from a reference one. We assume as reference distribution the empirical distribution function associated with the population of inliers. In particular, let

- n be the total number of outliers $|D^o|$,
- m be the number of inliers $|D^i|$,
- V^a be the set $\{v : v \in D^o[a] \vee v \in D^i[a]\}$ of the distinct values assumed by the attribute a in the datasets D^i and D^o,
- r be the number $|V^a|$ of distinct values in the domain of a, namely $V^a = \{v_1, \ldots, v_r\}$,
- f_i be the frequency of the value v_i in $D^o[a]$, that is $f_i = |\{t \in D^o : t[a] = v_i\}|$, and
- e_i be the frequency of the value v_i in $D^i[a]$, that is $e_i = |\{t \in D^i : t[a] = v_i\}|$.

The chi-square test relies on the chi-square statistic X^2, which is the sum of the squared difference between the observed frequencies and the expected ones, normalized on the value of the expected frequencies. Notice that the expected frequency associated with the value v_i can be obtained from the empirical distribution of the inliers as $e_i \frac{n}{m}$. Thus:

$$\mathrm{X}^2 = \sum_{i=1}^{r} \frac{\left(f_i - e_i\left(\frac{n}{m}\right)\right)^2}{e_i\left(\frac{n}{m}\right)} = \left(\sum_{i=1}^{r} \frac{f_i^2}{e_i\left(\frac{n}{m}\right)}\right) - n. \quad (1)$$

It is known that the X^2 statistic asymptotically approaches the χ^2 distribution with $r-1$ degrees of freedom, hence, the X^2 value is used to calculate a p-value[1] by comparing its value to the proper chi-squared distribution.

However, the direct application of the chi-square test to define a proper exceptionality score presents two main criticalities.

The first criticality is related to the correct computation of the formula in Equation (1) in the presence of expected frequencies e_i evaluating to zero. This problem can be solved by noticing that having some e_i equal to zero could be assimilated to the scenario in which the corresponding events have not been observed to occur at all due to the fact that the sample data is finite. In this case, underlying probabilities can be estimated by means of the *Laplace's rule of succession* [6]: when there is no other prior knowledge, the expected probability of a specified event is given by $\left(s + \frac{2}{r}\right)/(m+2)$, where s is the number of actual outcomes of the specified event in m actual samples, given the r events in total. Taking into account the above correction, in order to determine expected probabilities, the formula in Equation (1) can be replaced by the following one:

$$\mathrm{X}^2 = \left(\sum_{i=1}^{r} \frac{f_i^2}{\left(e_i + \frac{2}{r}\right)\left(\frac{n}{m+2}\right)}\right) - n. \quad (2)$$

By letting f denote $f_1, \ldots, f_r$ and e denote $e_1, \ldots, e_r$, the formula in Equation (2) is also denoted as $\mathrm{X}^2_e(f)$.

As for the second criticality, it must be noticed that when the expected values are small, the chi-square test (and also its alternatives, such as the G-test and others [6]) give inaccurate results. As a matter of fact, the significance value provided by the chi-square test is an approximation, because the distribution of the test statistic X^2 is only approximately equal to the theoretical chi-squared distribution χ^2. This approximation is inadequate when sample sizes are small, usually for $n \leq 1{,}000$, or the frequency counts are low. A common rule is that the generic frequency f_i should be equal to or greater than 5 in 80% of cases, and that there are no f_is equal to 0. Note that these conditions are not met in the scenario we consider here, since the outlier dataset consists of a small number n of objects, and, moreover, frequency counts f_i are also expected to be very small, with a sensible fraction of them being equal to 0.

In such a situation, in order to avoid inaccurate inference, alternative kinds of test, such as *exact* or *randomization tests*, are needed to be used. Exact tests are so called because the significance of the deviation from a null hypothesis is calculated exactly, rather than relying on an approximation that becomes exact only when the sample size grows to infinity. Here, a very natural alternative to the χ^2 test for small sample sizes

1. Recall that the p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis (in our case, that the observed distribution complies with the theoretical one) holds.

Function $\tau_a^{cat}(D^o, D^i)$

```
Set count to 0
for i ranging in 1..N do
    Randomly generate a bag B of n values from the set {v_1, ..., v_r}, with the probability p(v_i) of including v_i in B given by p(v_i) = (e_i + 2/r)/(m + 2)
    Let g = g_1, ..., g_r be the frequency counts in B associated with the elements v_i
    if X^2_e(g) ≥ X^2_e(f) then
        count = count + 1
p = count/N
return 1 − p
```

is the *Fisher's exact test* [7], [8]. This test works by calculating the probabilities of all possible combinations, that preserve both row and column totals, of integer numbers in an $r \times c$ contingency table (in our case, the table would consist of two columns, that is $c = 2$, one storing the values f_i, and the other one storing the values e_i), and then computing the sum of the probabilities of the combinations that are as extreme or more extreme than the observed data. However, as $r \cdot c$ gets larger or as the total sample size $n + m$ gets larger, the number of possible combinations increases dramatically, to the point that computing the test becomes infeasible. Unfortunately, this problem characterizes any other exact test, since these tests take their decision on a reference distribution obtained by calculating all possible values of the test statistic under rearrangements of the labels on the observed data, and, as such, have exponential cost both in the sample size and in the number of possible values in the domain.

When there are too many possible orderings of the data, in order to avoid complete enumeration, asymptotically equivalent exact tests, also called *randomization tests*, can be created by taking as reference distribution a relatively small sample of the whole set of combinations. Sometimes as few as $N = 400$ randomly generated combinations are sufficient to generate a reliable answer [6]. Thus, in order to obtain accurate inference on small samples, which is a peculiarity of outlier populations, we make use of a randomization test, as defined in the following.

The test works by randomly generating N bags B, each one consisting of n values taken from the domain $V^a = \{v_1, \ldots, v_r\}$, representing N random outcomes of the dataset $D^o[a]$. The probability $p(v_i)$ of inserting the value v_i in the bag B is equal to the probability of observing v_i under the null hypothesis, that is the frequency $(e_i + \frac{2}{r})/(m+2)$ of v_i in the inlier dataset. Let $g = g_1, \ldots, g_r$ be the frequency counts in B associated with the elements v_i. For each randomly generated bag B, Pearson's chi-square statistic $\mathrm{X}^2_e(g)$ (as defined in Equation (2)) is calculated. The fraction of these random outcomes that have a chi-square statistic equal to or greater than $\mathrm{X}^2_e(f)$, namely the value associated with the dataset actually observed, is the returned p-value.

The *exceptionality score* $\tau_a^{cat}(D^o, D^i)$ defined for categorical attributes is eventually computed as one minus the above computed p-value (hence, its value belongs to the interval $[0, 1]$). The computation of $\tau_a^{cat}(D^o, D^i)$ is illustrated in the figure above.

Recall that the lower the p-value, the less likely the result holds if the null hypothesis is true. Consequently, the alternative hypothesis is accepted (or, equivalently, the null hypothesis is rejected) if the p-value is less than the significance level α. We refer to the statistical literature (see, e.g., [8], [9]) according to which a well-founded value for α is 5% .

Thus a is an exceptional property for D^o w.r.t. D^i according to τ^{cat} at the significance level α if and only if $\tau_a^{cat}(D^o, D^a) > \theta_{cat}$, with $\theta_{cat} = 1 - \alpha$.

2.3 Numerical properties

Consider a numerical domain. A basic test used to assess whether a given distribution is suited to a sample, is the *Kolmogorov-Smirnov test*. This test uses as statistic the supremum of the absolute difference between the empirical distribution function F and the theoretical distribution function $\widehat{F}$, defined as:

$$\sup_x |F(x) - \widehat{F}(x)|.$$

The above statistic represents a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution. However, in practice, the Kolmogorov-Smirnov statistic requires relatively large number of data to properly reject the null hypothesis.

The *Cramér–von-Mises test* is an alternative to the Kolmogorov-Smirnov test. This test uses the integral of the squared difference between the empirical and the estimated distribution functions [3] as the statistic, defined as follows:

$$\omega^2 = \int_{-\infty}^{\infty} [F(x) - \widehat{F}(x)]^2 \, \mathrm{d}\widehat{F}(x) \tag{3}$$

In particular, let $S = \{x_1, x_2, \ldots, x_n\}$ be the set of observed values, listed in increasing order. Then, the following identity holds:

$$n\omega^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(\frac{2i-1}{2n} - \widehat{F}(x_i) \right)^2 . \tag{4}$$

The right term of Equation (4) will be denoted as $cvm_{\widehat{F}}(S)$.

If $cvm_{\widehat{F}}(S)$ is larger than a certain tabulated threshold value, then we can reject the hypothesis that the observed data comes from the distribution $\widehat{F}$. The following table reports the threshold values of the test for various sample sizes n (first row) and significance levels α (first column):

$\alpha \setminus n$	2	5	10	30	60	100
0.01	0.186	0.300	0.320	0.330	0.330	0.340
0.05	0.175	0.199	0.212	0.218	0.220	0.220

As already done for categorical domains, in this case we also employ as reference distribution $\widehat{F}$ the empirical distribution function associated with the population of inliers. Given a numerical attribute a, the reference distribution F_a associated with a is defined as:

$$F_a(x) = \frac{|\{D^i[a] : D^i[a] \leq x\}|}{|D^i|},$$

Then, the *exceptionality score* $\tau_a^{num}(D^o, D^i)$ defined for numerical attributes is computed as $cvm_{F_a}(D^o[a])$.

The rationale underlying the use of the *Cramér–von-Mises test* criterion as exceptionality score τ^{num} is that it is very appropriate for the task at hand, in which we compare a small population, composed of outliers, with a larger population. As a matter of fact, the criterion can be safely applied even when very small samples are available, as one including as few as just two observations [3].

2.4 Explanation-Property pair

Next, we provide the definition of exceptional explanation-property pair and of the associated mining problem.

For an attribute p, by τ_p we define either τ_p^{num} or τ_p^{cat}, depending on p being a numerical or categorical attribute, respectively.

Definition 2 (Exceptional Explanation-Property pair): Let D^o (outliers) and D^i (inliers) be two datasets on a set of attributes A, and let τ be an exceptionality score. A pair $\langle E, p\rangle$, where E is a set of conditions on A and p is an attribute of A not occurring in E, is an *exceptional explanation-property pair* in D^o w.r.t. D^i (or, simply, exceptional pair) if and only if the attribute p is an exceptional property for D^o_E w.r.t. D^i_E according to τ, that is, if $\tau_p(D^o_E, D^i_E) > \theta_\tau$. In this case the attribute p is said to be an *exceptional property* and the value $\tau_p(D^o_E, D^i_E)$ is called *exceptionality* (*score*) of p (with explanation E).

A desideratum of the explanation E is that it covers as many outliers as possible and a sensible fraction of inliers. Hence, we say that the exceptional pair is *representing*, if it holds that $|D^o_E| \geq \sigma^o|D^o|$ and $|D^i_E| \geq \sigma^i|D^i|$, where σ^o and σ^i are two suitable frequency thresholds.

Definition 3 (Exceptional Property Discovery Problem): Let D^o and D^i be two datasets on the same set of attributes. The Exceptional Property Discovery Problem is defined as follows: *find the representing exceptional explanation-property pairs in D^o w.r.t. D^i.*

3 Related Work

In this section, we compare our approach with some relevant literature somehow related with the presented work. For the sake of clarity, relevant papers are grouped into categories according to the specific task they are intended to solve, that are: *Outlier detection*, *Outlying property discovery*, *Emerging patterns*, *Contrast sets*, *Subgroup discovery*, *Rule-based classifiers*, and *Association rule mining*.

Outlier detection. Outlier detection is a well-known discovery problem [10]. We notice that the task considered here actually deals with outliers, but not in the classical sense of discovering them. In particular, approaches to outlier detection can be classified in supervised, semi-supervised, and unsupervised. Supervised methods exploit the availability of a labeled dataset, containing observations already labeled as normal and abnormal, in order to build a model of the normal class [11]. Semi-supervised methods assume that only normal examples are given and the goal is to find a rule partitioning the object space into an accepting region, containing the normal objects, and a rejecting region, containing all the other objects [12]. Unsupervised methods search for outliers in an unlabeled dataset by assigning to each object a score which reflects its degree of abnormality. These methods can be classified, in turn, as *distance-based* [13], [14], [15], *density-based* [16], *MDEF-based* [17], and *knowledge-based* [18], [19], [20], and others. It must be noticed that the problem addressed here is completely different from supervised and semi-supervised outlier detection, and, moreover, is to be considered orthogonal to the unsupervised outlier detection task. Indeed, in outlier detection, a set of observations is given in input and we are interested in discovering those observations (i.e., the outliers) that are mostly dissimilar from the remaining ones, while here the outliers (anomalous subpopulations) are given in input and we are interested in discovering the motivations underlying their abnormality.

Outlying property discovery. In [1] a data population is assumed to be given, characterized by a certain number of attributes, and the information is provided that one of the individuals in the data population is abnormal. In this context, the problem of discovering sets of attributes that account for the (a-priori stated) abnormality of such an individual is considered. Each subset of attributes is intended to represent a *property* of individuals. A property witnesses the abnormality of an object if the combination of values the object assumes on these attributes is very infrequent with respect to the overall distribution of the attribute values in the dataset, and this is measured by means of the so called *outlierness* function. Global and local properties are introduced. Global properties are subsets of attributes explaining the given abnormality with respect to the entire data population. With local ones, instead, two subsets of attributes are singled out. The first subset of attributes is used to select a subpopulation. In particular, only the individuals assuming on this subset of attributes the same value as the exceptional individual are selected. The second subset of attributes justifies the abnormality of the exceptional individual within the given sub-population.

As far as the differences with the approach [1] are

concerned, we note that the method proposed here is capable of discovering outlying properties associated with *groups* of anomalous individuals, while in [1] only *one single* anomaly can be compared with the normal population. In this respect, we note moreover that an approach supposedly consisting in finding the outlying properties of each anomaly of the group separately from the others, and then "merging" them to obtain properties valid for all the individuals of the group, would be weaker than the approach illustrated here and, in most cases, would produce no significant information. Indeed, the distribution on the whole set of anomalous individuals of the attribute values forming the properties would often get completely lost when the single anomalous observations are considered one at a time. This discussion will be substantiated in the section devoted to experimental results (see Section 5.2). We further point out that the criterion adopted here to measure the significance of the given property is completely different from the one employed in [1]. Indeed, the comparison of a single (abnormal) value with a potentially large number of (normal) values requires a very specialized test. As a matter of fact, [1] presented an *outlierness* score which is based on measuring how much the frequency of the combination of values assumed by that object on those attributes is rare as opposite to the frequencies associated with the other combinations of values assumed on the same attributes by the other objects in the population (and, in fact, in [1] the outlierness was shown to have some connections with the *Gini index* employed to measure the *heterogeneity* of a statistical distribution). In this respect, while this outlierness is appropriate for comparing one single value with a set of values, in the presence of two, even though unbalanced, distributions, a refined form of test must be employed in order to discover significant (and, often, subtler) characterizing properties. Finally, but not less relevant, since the approach pursued in [1] is based on measuring the frequency of the occurrence of some combination of values, it assumes that the attributes are categorical, while the approach presented here deals with both numerical and categorical attributes.

Emerging patterns and Contrast sets. Given two sub-populations, an *emerging pattern* (EP) [21], [22], [23] is a set of conditions whose support in one sub-population differs from its support in the other sub-population. A ρ-emerging pattern C is one in which the growth rate, namely the ratio between the supports of C in the two sub-populations, is greater than a user-provided threshold ρ. Several families of emerging patterns have been introduced in the literature [23]. For example, a pattern whose growth rate ρ is infinite is called *jumping* emerging pattern (JEP) [24]. It follows from the definition that a JEP holds for some individuals in one population and does not hold in the other. Conversely, *constrained* emerging patterns (CEP) [25] are the minimal set of conditions C such that the support of C in one sub-population is at least α and the support of C in the other one is at most β, where α and β are used-defined parameters. Contrast set mining [26], [27] was first introduced in [26], [28], and subsequently explored for classification purposes [29]. Given a population organized in two or more groups $G_1, \ldots, G_k$, the contrast set mining task looks for a set of conditions C, that are conjunctions of attribute-value pairs, also called *contrast sets*, whose support significantly changes across groups, namely the support of C is not independent of group membership. Moreover, for at least one pair of groups G_i and G_j, the difference of support of C in G_i and G_j must be at least as large as δ, where δ is a user-defined threshold.

Although there are some analogies between outlier-explanation pairs and emerging patterns/contrast sets, since both aim at discovering knowledge which is almost valid in one of the two sub-populations and almost invalid in the other one, relevant differences exist between the two kinds of knowledge. First of all, the form of the mined knowledge is different, since an explanation-property pair consists in a set of conditions (the explanation) and in one attribute (the property), while emerging patterns/contrast sets take the form of sets of conditions. Thus, the latter techniques do not single out properties characterizing differences across sub-populations. Notably, differently from emerging patterns, our approach deals with both categorical and numerical attributes. In particular, no discretization step is needed in order to treat numerical attributes as properties. Furthermore, the significance test exploited by emerging patterns/contrast sets is based on the support of the pattern in the sub-populations. Importantly, our exceptionality property cannot be mapped to an attribute-value pair (and, hence, simply incorporated in an emerging pattern), since the exceptionality score compares the distributions associated with the exceptional attribute in the two sub-populations. To conclude, emerging patterns/contrast sets capture knowledge characterizing a population in a *global* sense, since they are based on the notion of support which is related to the absolute frequency of the itemset. Conversely, the knowledge mined by means of the explanation-property pairs characterizes the population in a *local* sense. Indeed, the explanation selects a portion of the sub-populations in which the difference in the "behavior" of the property has a strong evidence.

Subgroup discovery. The *Subgroup Discovery Task* (SDT) aims at finding an interesting subgroup of objects with common characteristics with respect to a given attribute-value pair, called *target variable* [23], [27]. It was introduced in [27], [30] for categorical domains and, recently, extended to numerical domains [31]. The SDT outputs a subgroup of individuals, identified as those individuals of the population satisfying a set of conditions, whose behavior on the target attribute-value pair is different from the behavior of the population taken as a whole.

Even if there are some relationships with the task addressed here, the SDT presents relevant differences with it. In SDT the behavior of the subgroup is compared

with the whole population and SDT does not deal with a-priori provided sub-populations. One may think that the SDT can be exploited to solve the exceptional property discovery task by pursuing the following approach: First, merge the two sub-populations in one population, then encode in a binary attribute the information about the sub-populations and, finally, perform the subgroup discovery task. Nevertheless, this approach is not suited for the task presented here. Indeed, first of all, one sub-population is compared with the whole population and not with the other sub-population; second, no property (attribute) distinguishing a sub-population from the other one would be singled out, while our task is precisely designed to find such properties (possibly with an associated explanation). To conclude, our tests of significance of the explanation-property pairs have been designed to handle unbalanced sub-populations, while it is not the case for SDT.

Rule-based (rare class) classifiers and Association rule mining. The task addressed here is substantially different also from the rule-based classification task [32], even when the classifier is trained to predict rare classes [33]. Rules induced by a rule-based classifier are of the form $X \rightarrow c$, where X is a set of conditions and c is one of the classes. The goal is to exploit induced rules in order to predict the class of unclassified objects.

As for rule based-classifiers, while a rule $X \rightarrow c$ encodes the information that the class of the individuals satisfying the set of conditions X is (almost always) c, our goal is to single out attributes (properties) showing different behavior across classes. Thus, the criteria for measuring the significance of the induced knowledge are sharply different. As for association rules, association rule miners [34] search for rules of the form $X \rightarrow Y$, where X and Y are sets of attribute-value pairs, standing for *almost all individuals satisfying X satisfy also Y*. Their purpose is to find rules which are clearly represented in the dataset at hand, but they do not distinguish between sub-populations. Moreover, even enforcing the mining algorithm in employing only a binary variable encoding the sub-population information as the rule head, the results would be substantially different from those our approach allows to obtain, inasmuch as this is precisely the technique employed by rule-based classifiers in order to mine their rules.

4 ALGORITHM

In this section we present the EXPREX algorithm (meaning EXceptional PRoperty EXtractor) which solves the Exceptional Property Discovery Problem introduced in the previous section. In particular, Section 4.1 describes the algorithm, while Section 4.2 analyzes its cost.

4.1 EXPREX algorithm

In order to solve the exceptional property discovery problem the EXPREX algorithm performs, for each attribute p of the dataset, the following two main steps:

Function *GenerateBaseConditions(D^o, D^i, a, σ^o, σ^i)*

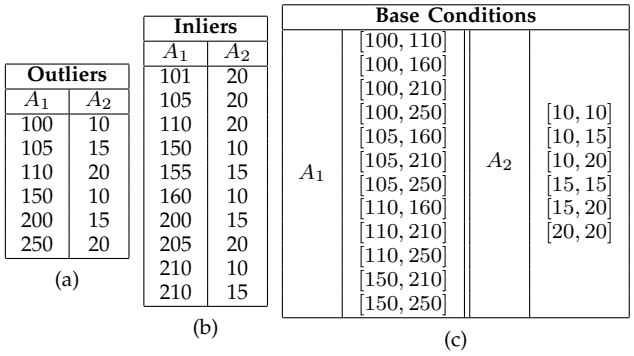
Input: the outlier dataset D^o
the inlier dataset D^i
the attribute a
the outlier frequency threshold σ^o
the inlier frequency threshold σ^i
Output: the set of *base* conditions C^a for the attribute a

set $\mathcal{I}^a$ to $\emptyset$
if *a is a categorical attribute* **then**
  foreach v *in* $(D^o[a] \cap D^i[a])$ **do**
    if $|D^o_{a=v}| \geq \sigma^o|D^o|$ *and* $|D^i_{a=v}| \geq \sigma^i|D^i|$ **then**
      add v to $\mathcal{I}^a$
else // *a is a numeric attribute*
  set Out to the ordered set of values $D^o[a] \cup \{+\infty\}$
  set In to the ordered set of values $D^i[a]$
  set V to the ordered set of values $Out \cup In$
  set l' to the first value in Out
  set l to minimum between l' and the first value in In
  while $[l, +\infty)$ *captures at least* $\sigma^o|D^o|$ *outliers and* $\sigma^i|D^i|$ *inliers* **do**
    set u' to the $\sigma^o|D^o|$-th value of Out greater than l'
    repeat
      set u'' to the succeeding value of u' in Out
      set u to the maximum value in V belonging to $[u', u'')$
      if $[l, u]$ *selects at least* $\sigma^i|D^i|$ *inliers* **then**
        add $[l, u]$ to $\mathcal{I}^a$
      set u' to u''
    until u' *is lower than* $+\infty$
    set l to the lowest value in V strictly greater than l'
    set l' to the value succeeding l' in Out
set C^a to $\emptyset$
for i *ranging in* $1..|\mathcal{I}^a|$ **do**
  add the condition $(a \in \mathcal{I}^a_i)$ to C^a
return C^a

Function *CombineBaseConditions(D^o, D^i, p, B, σ^o, σ^i)*

Input: the outlier dataset D^o
the inlier dataset D^i
the current property p
the total set of base conditions B
the outlier frequency threshold σ^o
the inlier frequency threshold σ^i
Output: the exceptional explanation-property pairs with property p

set S to $\emptyset$
if $\tau(\emptyset, p)$ *is greater than* θ_τ **then**
  add $\langle \emptyset, p \rangle$ to S
set $Curr$ to $\{\emptyset\}$
repeat
  foreach *condition C in $Curr$* **do**
    foreach *condition b in B associated with an attribute not occurring in C* **do**
      if $C \cup \{b\}$ *is not in* $Next$ **then**
        set C' to $C \cup \{b\}$
        if *all the subsets of C' of size $|C|$ are in $Curr$ and both* $|D^o_{C'}| \geq \sigma^o \cdot |D^o|$ *and* $|D^i_{C'}| \geq \sigma^i \cdot |D^i|$ **then**
          add C' to $Next$
          if $\tau(C', p)$ *is greater than* θ_τ **then**
            add $\langle C', p \rangle$ to S
  set $Curr$ to $Next$
until $Curr$ *is empty*
return S



Outliers	
A_1	A_2
100	10
105	15
110	20
150	10
200	15
250	20

(a)

Inliers	
A_1	A_2
101	20
105	20
110	20
150	10
155	15
160	10
200	15
205	20
210	10
210	15

(b)

Base Conditions			
A_1	$[100, 110]$	A_2	$[10, 10]$
	$[100, 160]$		$[10, 15]$
	$[100, 210]$		$[10, 20]$
	$[100, 250]$		$[15, 15]$
	$[105, 160]$		$[15, 20]$
	$[105, 210]$		$[20, 20]$
	$[105, 250]$		
	$[110, 160]$		
	$[110, 210]$		
	$[110, 250]$		
	$[150, 210]$		
	$[150, 250]$		

(c)

Fig. 1: Example of base condition generation

Algorithm 1: EXPREX algorithm

Input: the outlier dataset D^o and
the inlier dataset D^i
the outlier frequency threshold σ^o
the inlier frequency threshold σ^i
Output: the exceptional explanation-property pairs

1: let A be the set of attributes of D^o and D^i
2: set R to $\emptyset$ *// the set storing the exceptional explanation-property pairs*
3: **foreach** $a \in A$ **do**
// base condition generation step
4: $C^a = GenerateBaseConditions(D^o, D^i, a, \sigma^o, \sigma^i)$
5: **foreach** $p \in A$ **do**
// base condition combination step
6: let B be the whole set of conditions $\bigcup_{a \in (A \setminus \{p\})} C^a$
7: $R = R \cup CombineBaseConditions(D^o, D^i, p, B, \sigma^o, \sigma^i)$
8: **return** R

- *Base condition generation step*: a set of base conditions is associated with each attribute different from p;
- *Base condition combination step*: the base conditions associated with distinct attributes are combined together to discover the representing explanation-property pairs.

Next we illustrate the two aforementioned steps.

Base condition generation step. This step is taken care of by the function *GenerateBaseConditions*.

A set of conditions C_a, called *base conditions*, is associated with each attribute a. These conditions are then combined during the subsequent *base condition combination step* in order to form relevant explanations.

The base conditions associated with categorical attributes a are all the conditions of the form $a = v$, where v is a value belonging to the domain $D^o[a] \cap D^i[a]$, and both $|D^o_{a=v}| \geq \sigma^o|D^o|$ and $|D^i_{a=v}| \geq \sigma^i|D^i|$ hold (lines 3–5).

As far as numerical attributes a are concerned, it must be noticed that the size of the set of conditions of the form $a \in [v, u]$, where both v and u are real numbers in the *actual domain* $D^o[a] \cup D^i[a]$ of a, is quadratic in the number of dataset objects.

Thus, in order to reduce the large search space associated with this kind of conditions, the EXPREX algorithm adopts a strategy consisting in selecting the relevant subsets of the overall set of conditions.

In particular, the *base intervals* for numerical attributes are the intervals $I = [l, u]$, with l and u belonging to the actual domain of a, satisfying the following points:

1) $|D^o_{a \in I}|$ is at least $\lceil \sigma^o \cdot |D^o| \rceil$,
2) $|D^i_{a \in I}|$ is at least $\lceil \sigma^i \cdot |D^i| \rceil$,
3) there is no interval $[l', u'] \supset [l, u]$, with l' and u' belonging to the actual domain of a, capturing exactly the same outliers as $[l, u]$ and satisfying both points 1 and 2.

Intuitively, base intervals are the largest intervals I capturing a certain set of outliers (by points 3) which satisfies the frequency constraints on the number of outliers (by point 1) and inliers (by point 2) selected by the condition $a \in I$.

In order to compute the base conditions for a numeric attribute a, the function starts by sorting the set of values assumed by the dataset objects on the attribute a and storing them in the set V (lines 7-9). Then, the function builds the base conditions, starting from the minimal intervals which can be defined on the outlier set.

In particular, let $[l', u']$ be an interval capturing exactly $\sigma^o|D^o|$ outliers, with l' and u' being two values in $D^o[a]$ (see line 13). The function, in order to satisfy point 3, determines the interval $[l, u] \supseteq [l', u']$, where $[l, u]$ is the largest interval capturing exactly the same outliers as $[l', u']$ and both l and u belong to V (lines 11 and 21). In particular, the lower bound l is set either at line 11 (only the first time) or 21 (all the other times), while the upper bound u is set at line 16.

The set $[l, u]$ satisfies both points 1 and 3 by construction. If it also satisfies point 2, then it is added to the set of base conditions (line 17-18).

Then, the function fixes the lower bound l and considers a new upperbound u' generating a new interval capturing exactly one outlier more than $[l, u]$, and reiterates the previously described operations (lines 14–20).

Once all the intervals with lowerbound l have been generated, the function considers the next minimal interval capturing exactly $\sigma^o|D^o|$ outliers (lines 21–22).

The cost of this function will be discussed in Section 4.2.

Figure 1 reports an example to illustrate the function *GenerateBaseConditions*. Assume to set the support thresholds σ^o and σ^i to 0.3, then at least 2 outliers and 3 inliers should be captured by the explanations. Consider the attribute A_1. The first interval $[l, u]$ containing at least 2 outliers is $[l, u] = [100, 105]$ ($l = \min\{100, 101\}$, $u' = 105$, $u'' = 110$, and $u = \max\{v : v \in V \cap [u', u'')\}$, that is $u = \max\{v : v \in [105, 110)\} = 105$). Since the interval $[l, u] = [100, 105]$ contains only 2 inliers, it has to be enlarged. So u' is set to 110, u'' to 150 and $u = \max\{v : v \in V \cap [u', u'')\} = 110$. Now, the interval $[l, u] = [100, 110]$ contains at least 3 inliers and hence it is added to the set of base conditions. Then, u' is set to $u'' = 150$ and the next iteration starts. The value u'' is set to 200, u to $\max\{v : v \in V \cap [u', u'')\} = 160$, and the interval $[100, 160]$ is added to the set of base

conditions. The function continues in a similar manner. Figure 1(c) shows the set of base conditions computed by the function *GenerateBaseConditions* on the example database.

Base condition combination step. This step is taken care of by the function *CombineBaseConditions*. It tries to combine base conditions associated with different attributes, and follows an a-priori like strategy [34] to find the set of conditions E such that $|D^o_E|$ ($|D^i_E|$, resp.) is not smaller than $\sigma^o \cdot |D^o|$ ($\sigma^i \cdot |D^i|$, resp.). The pairs $\langle E, p\rangle$ having exceptionality score greater than the threshold θ_τ are collected and then stored in the set R which maintains all the exceptional explanation-property pairs discovered so far.

Consider, again, the example reported in Figure 1. Since the datasets have two attributes, there are two properties to take into account. Then, the algorithm calls twice the function *CombineBaseConditions*. When the function *CombineBaseConditions* is called on the property $p = A_1$ ($p = A_2$, resp.), the set B consists in the set of basic conditions generated for the attribute A_2 (A_1, resp.).

Consider the function *CombineBaseConditions* called on the property $p = A_2$. First of all, lines 2-3 test if the property A_2 with empty explanation is exceptional. Next, the exceptionality of A_2 is tested when any base condition C on A_1 is used as explanation. If the explanation-property pair $\langle C, A_2\rangle$ is exceptional, it is added to the set S to be returned in the result. An example of explanation-property pair is $\langle A_1 \in [100, 110], A_2\rangle$. Indeed, while the values assumed on A_2 by the set of outliers selected by the explanation are uniformly distributed in the range $[10, 20]$, the values assumed by the inliers selected by the explanation are likely to come from a very different distribution, since they are all equal to 20. Note that just one iteration of the cycle between lines 5 and 15 is performed. Indeed, after the first iteration, $Curr$ contains all and only the base conditions on A_1 and there is no condition $b \in B$ associated with an attribute not occurring in a condition in $Curr$.

4.2 Cost of the algorithm

In this section, the temporal cost of the EXPREX algorithm is analyzed. Let D^o and D^i be the outlier and the inlier dataset, respectively, both defined on the set of attributes A. Let d be the size of A and let n (m, resp.) be the total number of outliers (inliers, resp.), namely $n = |D^o|$ and $m = |D^i|$.

Consider the *GenerateBaseConditions* function. Given an attribute a, the cost of computing the set of base conditions for a is $O(n+m)$, if a is a categorical attribute. Conversely, if a is a numeric attribute, the cost required by this function is $O(n^2 + m \log m)$. Indeed, the function first sorts the sets $D^o[a]$ and $D^i[a]$, an operation requiring time $O(n \log n + m \log m)$. The outer cycle (lines 12–22) of EXPREX is performed at most $|D^o|$ times, since, at each iteration, l assumes a different value in $D^o[a]$. The inner cycle (lines 14–20) is also performed at most $|D^o|$ times, since, at each iteration, u' assumes a different value in $D^o[a]$. As for the other operations, they all require $O(1)$ time, since the array Out and V are sorted.

Consider, now, the *CombineBaseConditions* function. Assume that, for each attribute a, there are at most h base conditions. Essentially, this function evaluates the exceptionality score for each set C of conditions which can be generated from the overall set of base conditions. Let $c(\tau)$ denote the cost of evaluating the exceptionality score for a given condition. Since, for each attribute a, at most one of the h base conditions defined on a can occur in C, the number of sets of base conditions is $(h+1)^{d-1}$, where d is the number of attributes. Thus, the *CombineBaseConditions* function costs $O((h+1)^{d-1} \cdot c(\tau))$. Therefore, the overall cost of the EXPREX algorithm is $O(d \cdot (h+1)^{d-1} \cdot c(\tau))$.

Before concluding, it must be noticed that this is a worst case cost and that, in practice, the number of exceptionality score evaluations is much smaller due to the strategy adopted to explore the search space.

5 Experiments

In this section, we present experiments conducted by using the EXPREX algorithm. The experiments are designed as follows. First, we consider some real datasets, including both numerical and categorical domains, in order to assess the capability of the approach in mining interesting knowledge (Section 5.1). Then, we compare our technique with the outlying property discovery technique described in [1], in order to point out differences and to show that the approach we present here is more powerful in characterizing groups of outliers (Section 5.2). Finally, we discuss results obtained on the longevous individual characterization scenario already described in the Introduction (Section 5.3). If not otherwise stated, we ran the EXPREX algorithm with signficance level $\alpha = 0.05$.

5.1 Experiments on real data

In order to show the behavior of the algorithm, we first considered four real datasets from the *UCI Machine Learning Repository*, that are *Abalone*, *Ecoli*, *Parkinsons*, and *SPECT Heart* [35]. In the following, for some of the *top* (i.e. those scoring the largest value of exceptionality score) explanation-property pairs $\langle E^*, p^*\rangle$ discovered by the algorithm, we will compare the distribution of $D^o_{E^*}[p^*]$ and $D^i_{E^*}[p^*]$ with that of $D^o[p^*]$ and $D^i[p^*]$, in order to make the quality of the separation between outliers and inliers thus obtained intelligible.

Abalone. The *Abalone* dataset contains information about a population of abalones (haliotis) of the Tasmania. There are seven numerical attributes, that are *length* (mm), *diameter* (mm), *height* (mm), *whole weight* (grams), *shucked weight* (grams), *viscera weight* (grams), *shell weight* (grams). The ranges of these attributes are scaled by

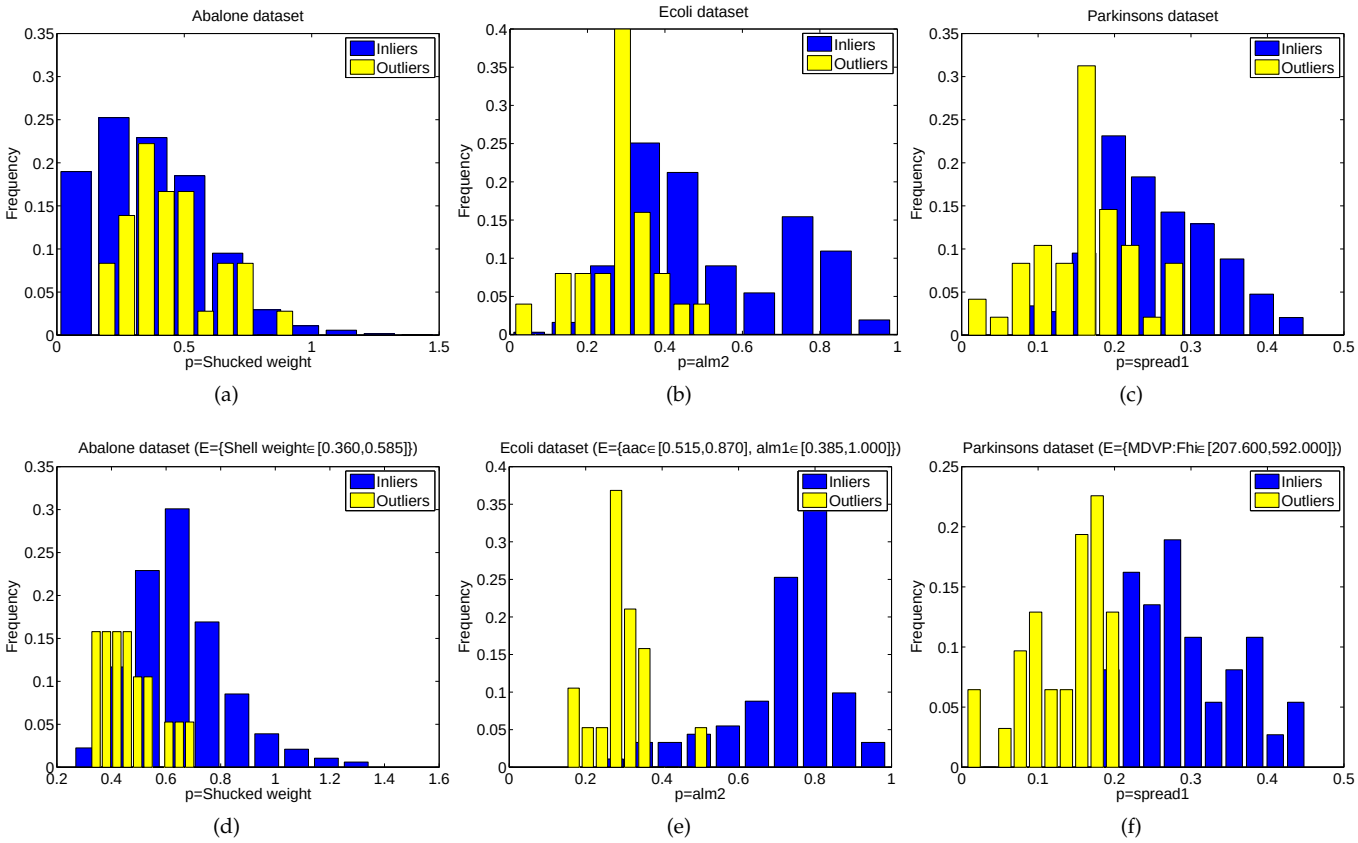


Fig. 2: Experimental results on real-life data.

dividing the values by 200. We selected as the outlier dataset D^o_{ab}, the thirty-six abalones having more than 20 rings on the shell (this number basically provides the animal age in years, that can be obtained by adding to it the constant 1.5) while the 4,141 remaining ones form the inlier dataset D^i_{ab}.

Figure 2d shows the top explanation-property pair of this data set. In particular, the outlying property p_{ab} is *shucked weight*, that is the weight of abalone meat, and the explanation E_{ab} includes the single condition *shell weight* $\in [0.360, 0.585]$, that is the weight after drying the abalone. This pair has exceptionality score $\tau = 2.48$, while the condition E_{ab} selects 19 (53%) outliers and 668 inliers (16%). Figure 2a shows the distribution of the attribute *shucked weight* in the whole populations of inliers and outliers. Without considering the explanation E_{ab} (that is to say, if E is empty), the attribute *shucked weight* has exceptionality score $\tau = 0.88$.

It is clear indeed that, while the distribution of *shucked weight* in the two sub-populations has some commonalities (Figure 2a), if the attention is restricted to the instances satisfying condition E_{ab} then the two distributions are markedly different (Figure 2d). Intuitively, the knowledge mined can be summarized as follows: among the abalones having medium/high weighted shells, the older ones are characterized by having less meat.

Ecoli. The *Ecoli* dataset contains information concerning the Escherichia coli bacterium, which is a gram-negative bacterium commonly found in the lower intestine of endothermic organisms. There are seven numeric attributes here, namely *mcg*, *gvh*, *lip*, *chg*, *aac*, *alm1*, and *alm2* (refer to [35] for details). The class attribute represents the localization site and we selected as the outlier dataset D^o_{ec} the twenty-five lying in the outer membrane (*om* and *omL* classes) while the 311 remaining ones form the inlier dataset D^i_{ec} (with location in the cytoplasm, perisplasm, and inner membrane).

Figure 2e shows the top explanation-property pair for this data set. In particular, the outlying property p_{ec} is *alm2* (score of ALOM program after excluding putative cleavable signal regions from the sequence) and the explanation E_{ec} includes the two conditions *acc* $\in [0.515, 0.870]$ (score of discriminant analysis of the amino acid content of outer membrane and periplasmic proteins) and *alm1* $\in [0.385, 1.000]$ (score of the ALOM membrane spanning region prediction program). The pair $\langle E_{ec}, p_{ec} \rangle$ has exceptionality score $\tau = 6.31$, while the condition E_{ec} selects 19 outliers (76%) and 91 inliers (29%).

Figure 2b shows the distribution of the attribute *alm2* in the whole populations of inliers and outliers. Without considering the explanation E_{ec} the attribute *alm2* has exceptionality score $\tau = 4.10$. In this case, the explanation-property pair we found is able to single out two almost

well-separated distributions.

Parkinsons. The *Parkinsons* dataset contains twenty two biomedical voice measures from 31 people, 23 with Parkinson's disease. Each attribute records a particular voice measure, and each instance corresponds to one of 195 voice recordings from these individuals. The class attribute represents the health status, and we selected as the outlier dataset D^o_{pk} the 48 voice recordings of healthy persons, while the 147 remaining ones form the inlier dataset D^i_{pk}.

Figure 2f shows the top explanation-property pair for this data set. In particular, the outlying property p_{pk} is *spread1* (a nonlinear measure of fundamental frequency variation) and the explanation E_{pk} only includes the condition *MDVP:Fhi* $\in [207.6, 592.0]$ (maximum vocal fundamental frequency in Hertz). The pair $\langle E_{pk}, p_{pk} \rangle$ has exceptionality score $\tau = 10.13$, while the condition E_{pk} selects 32 outliers (67%) and 37 inliers (25%).

Figure 2c shows the distribution of the attribute *spread1* in the whole populations of inliers and outliers. Without considering the explanation, the attribute *spread1* has exceptionality score $\tau = 5.83$. Also in this case, the selected explanation-property pair is able to single out two almost well-separated distributions.

SPECT Heart. The *SPECT* dataset contains information about diagnosis from cardiac Single Proton Emission Computed Tomography (SPECT) images. The dataset consists in 172 inliers and 15 outliers. The input SPECT images were processed and summarized in the dataset as 22 binary attributes, numbered from F_1 to F_{22}. Next, we comment on some outstanding explanation-property pairs. Remarkably, the properties F_{13} and F_{22} have exceptionality score 1.0 with empty explanation. The following tables show the distribution of the values associated with these two attributes.

	D^o	D^i
0	13	70
1	2	102

Property F_{13}.

	D^o	D^i
0	14	89
1	1	83

Property F_{22}.

Another top explanation-property pair reported by EXPREX is the attribute F_9 with explanation $E = \{F_4 = 0, F_6 = 0, F_{10} = 0\}$. The distribution of the attribute F_9 with empty explanation is approximatively the same for both the inlier sub-population and the outlier one. In fact, about 63% of the inliers and 66% of the outliers assume value 0 on the attribute F_9, as reported in the following table.

	D^o	D^i
0	10	109
1	5	63

F_9 without explanation.

	D^o	D^i
0	8	54
1	3	1

F_9 with explanation.

The value of the exceptionality score for the property F_9 without explanation is $\tau = 0.191$, which is much lower than the significativity threshold $\alpha = 0.95$. The explanation E selects 73% of the outliers (11 objects) and the 32% of inliers (55 objects). It must be noted that the distribution of the attribute F_9 on the selected individuals is substantially different. In fact, above 98% of inliers (54 objects) and about 72% of outliers (8 objects) assume value 0 on the attribute F_9. Also in this case, the value of the exceptionality is $\tau = 1.0$, which is larger than the significativity threshold $\alpha = 0.95$.

5.2 Comparison with outlying property discovery

In this section we compare the EXPREX algorithm with the *outlying property discovery* technique (OPD, for short) presented in [1]. The OPD works only on datasets consisting of categorical attributes. We thus compared the two techniques on the *SPECT Heart* dataset which has been illustrated in the previous section.

We recall that the ODP technique takes in input the inlier dataset and one single outlier object o, and returns a set of *outlying explanation-property pairs* $(\mathbf{E}, \mathbf{S})$, where both $\mathbf{E}$ and $\mathbf{S}$ are set of attributes. In particular, the set $\mathbf{E}$ is the analogous of the concept of explanation E adopted here, in that it implicitly defines the condition $E = \{a = o[a] : a \in \mathbf{E}\}$ aimed at singling out a portion of the inliers, while $\mathbf{S}$ plays the role of our property p, in that it consists of attributes on which the outlier object o assumes values to be considered exceptional within the sub-population identified by the condition E. Since the algorithm of [1] takes as input just one abnormal individual, we ran the OPD solving algorithm separately on each outlier and collected the set $\mathbf{T}$ consisting of the ten top ranked pairs returned by OPD on each outlier. Then, we analysed the explanation-property pairs in the set $\mathbf{T}$ by computing the exceptionality score τ introduced here on these pairs, in order to study how they behave in characterizing the whole outlier sub-population.

We executed the OPD algorithm with *explanation threshold* $\sigma = 0.3$ (denoting that at least 30% of the inliers have to satisfy the explanation). The top ranked explanation-property pairs returned by the ODP solving algorithm are shown in Table 1. Each row of the table is associated with one of the fifteen outlier objects and reports the pair scoring the largest outlierness value. It must be noticed that by virtue of the implied projection, in the outlier sub-population there are seven identical objects. These objects are identified in the table by means of a star located next to the object identifier. Clearly, the explanation-property pairs associated with all these objects are the same. Consider now the top ranked pair associated with the seven identical outliers. We note that its exceptionality score is far above the exceptionality threshold $\theta = 0.95$ (corresponding to significance level $\alpha = 0.05$ for categorical attributes) and, moreover, that this pair scores the largest value of exceptionality score τ associated with the top pairs reported by the OPD algorithm. The good behavior of this pair as far as the exceptionality score is concerned can be justified by noticing that it has been induced several times by the

ID	Property, p	Explanation, E	$\lvert D_E^o \rvert\%$	τ
1^*	F_{13}	$(F_{19}=0)$, $(F_{16}=0)$, $(F_{20}=0)$, $(F_1=0)$	0.60 (9)	0.9988
2	F_4	$(F_{18}=0)$, $(F_{17}=0)$, $(F_{15}=0)$, $(F_{19}=0)$, $(F_7=0)$, $(F_{16}=0)$	0.67 (10)	0.3768
3^*	F_{13}	$(F_{19}=0)$, $(F_{16}=0)$, $(F_{20}=0)$, $(F_1=0)$	0.60 (9)	0.9988
4	F_{19}	$(F_{18}=0)$, $(F_{17}=0)$, $(F_{15}=0)$, $(F_6=0)$, $(F_{10}=0)$	0.80 (12)	0.6538
5	F_9	$(F_{18}=0)$, $(F_{17}=0)$, $(F_{11}=0)$, $(F_{15}=0)$, $(F_4=0)$, $(F_{20}=0)$	0.73 (11)	0.9540
6^*	F_{13}	$(F_{19}=0)$, $(F_{16}=0)$, $(F_{20}=0)$, $(F_1=0)$	0.60 (9)	0.9988
7^*	F_{13}	$(F_{19}=0)$, $(F_{16}=0)$, $(F_{20}=0)$, $(F_1=0)$	0.60 (9)	0.9988
8^*	F_{13}	$(F_{19}=0)$, $(F_{16}=0)$, $(F_{20}=0)$, $(F_1=0)$	0.60 (9)	0.9988
9	F_{19}	$(F_{18}=0)$, $(F_{17}=0)$, $(F_{15}=0)$, $(F_1=0)$	0.80 (12)	0.8708
10	F_9	$(F_{18}=0)$, $(F_{17}=0)$, $(F_{15}=0)$, $(F_{19}=0)$, $(F_6=0)$, $(F_{14}=0)$	0.53 (8)	0.0000
11^*	F_{13}	$(F_{19}=0)$, $(F_{16}=0)$, $(F_{20}=0)$, $(F_1=0)$	0.60 (9)	0.9988
12^*	F_{13}	$(F_{19}=0)$, $(F_{16}=0)$, $(F_{20}=0)$, $(F_1=0)$	0.60 (9)	0.9988
13	F_{12}	$(F_{18}=0)$, $(F_7=0)$, $(F_{21}=0)$, $(F_3=0)$	0.73 (11)	0.6692
14	F_{15}	$(F_{18}=0)$, $(F_{17}=0)$, $(F_{19}=0)$, $(F_9=0)$	0.60 (9)	0.8782
15	F_9	$(F_{18}=0)$, $(F_{17}=0)$, $(F_{15}=0)$, $(F_4=0)$, $(F_{10}=0)$	0.67 (10)	0.9834

TABLE 1: Top ranked explanation-property pairs induced by the technique introduced in [1].

OPD technique, each time on one of the seven identical outlier objects (about 50% of the outliers). Thus, it can be said that this pair characterizes a large portion of the outliers, which is precisely the task we pursue here. Nonetheless, it has been shown in the previous section that the property F_{13} has exceptionality score 1.0 with empty explanation, hence slightly greater than the score associated with the above discussed pair. Thus, although this pair is significant, it can be just considered, loosely speaking, sub-optimal for our task. As for the top ranked rule associated with the other outliers, we note that in some cases the value of exceptionality score τ is larger than the exceptionality threshold $\theta = 0.95$ (see outliers with ID 5 and 15), while in all the other cases (see outliers with ID 2, 4, 9, 10, 13, and 14) the value of τ is lower than θ and, sometimes, very small.

From the previous analysis, it is clear that there is no guarantee on the value of the exceptionality score associated with pairs returned by the OPD technique. This can be explained by noting that the OPD technique is interested in maximizing the outlierness score, which depends only on the value assumed by the outlier object and does not take into account the values characterizing the other outliers. Also, attempting to combine (according to some strategy which should be properly defined) the pairs induced by the OPD technique on different outliers, would certainly result in a weaker approach than the one we present here. For example, just notice that in this experiment, no top outlying explanation-property pair is shared by different outliers and, hence, taking the intersection of the top pairs would result in an empty set. Thus, it can be concluded that the pairs returned by [1] are not suitable for characterizing an outlier sub-population. More importantly, it must be noticed that there are outstanding exceptional pairs that the OPD technique is not able to induce. As a notable example, consider the property F_{22} with empty explanation discussed in the previous section. This property has not been reported among the outlying pairs. In order to understand why OPD fails in recognizing such an exceptional property, observe the distribution of the values of the attribute F_{22} reported in the previous section. Within the inlier sub-population, the values 0 and 1 exhibit about the same frequency, that is about fifty percent. In this case, it is clear that if each outlier is taken into account separately, there is no way for them to show an exceptional value (that is, either 0 or 1) on the attribute F_{22}. On the contrary, by taking into account the whole outlier sub-population, it appears that the distribution of the attribute F_{22} is unexpectedly unbalanced.

5.3 Longevous individual characterization

The dataset *Longevous*, already discussed in the Introduction, consists of genotype information taken from the DNA of 972 unrelated subjects of various ages, ranging from 18 to 106 years. All the subjects were sampled from a genetically homogeneous population (Calabria, Southern Italy). Data were collected through a recruitment campaign. In particular, the subjects over 80 years old were identified through the birth registers of the municipalities of Calabria. The individuals aged from 18 to 60 years were sampled from the students and the staff of the University of Calabria. Finally, the individuals from 60 to 80 years were recruited among people visiting thermal baths in the area and the Academy of the Elderly. For each individual, the dataset stores information concerning ten polymorphic genetic loci. A genetic locus is the position of a gene or of any other significant sequence in a chromosome. A genetic locus is composed by a sequence of nucleotides, and each variant of this sequence is called *allele*. If there is more than one allele with frequency larger than 1% for a locus in a population, then the locus is called *polymorphic*. The polymorphic genetic loci considered in the dataset are: "APOA1", "APOA4", "APOB", "APOE", "HSP70-1", "HSP90α", "HSP90β", "SIRT3", "TH" and "mtDNA" due to their functional effects which are documented to be related to longevity (see [4] and the related literature cited therein). Summarizing, the dataset consists in 972 individuals and 10 binary attributes. The attribute values have been determined by the biologists on the basis of biological considerations about dominant alleles. Specifically, for each genetic locus, the value 1 is assigned to

a particular genotype, called *relevant genotype*, while the value 0 to the remaining ones. Moreover, according to the age ranges, the inlier dataset is composed of 875 subjects (individuals at most 100 years old), while the outlier dataset of 97 outliers (individuals more than 100 years old).

As (partially) discussed in the Introduction, the algorithm singled out the following exceptional pairs:

(i) $p =$ APOE and $E = \{\text{SEX} = \text{“}F\text{”}\}$, with exceptionality score $\tau = 0.995$;
(ii) $p =$ HSP90-β and $E = \{\text{SEX} = \text{“}F\text{”}\}$, with exceptionality score $\tau = 0.972$;
(iii) $p =$ APOE and $E = \{\text{SEX} = \text{“}M\text{”}\}$, with exceptionality score $\tau = 0.970$;
(iv) $p =$ SIRT3 and $E = \{\text{SEX} = \text{“}M\text{”}\}$, with exceptionality score $\tau = 0.995$;
(v) $p =$ APOA1 and $E = \{\text{SEX} = \text{“}M\text{”}\}$, with exceptionality score $\tau = 0.992$;

which agree with the results of the biological study presented in [4]. Other than the above shown exceptional pairs, EXPREX also mined other pairs which show a clear influence on the longevity. We discuss some of them in the following.

The property $p =$ APOA4 with explanation $E = \{\text{APOB} = 0, \text{APOE} = 0, \text{HSP90-}\beta = 1\}$ has exceptionality score $\tau = 0.971$ and involves 34 outliers and 300 inliers. The distributions of the property APOA4 with empty explanation and with the explanation E mined by the algorithm are reported in the following tables:

	D^o	D^i
0	64	612
1	33	263

“APOA4” without explanation.

	D^o	D^i
0	18	210
1	16	90

“APOA4” with explanation.

It is clear that if no explanation is considered, the property APOA4 is not exceptional. Indeed, since the two distributions are almost identical (the percentage of subjects assuming value 1 on the attribute APOA4 is about 30% for both outliers and inliers), the score τ is equal to 0.5686, thus far below the threshold 0.95. Conversely, by considering the sub-population selected by the explanation E, the inliers assuming value 1 on APOA4 are the 30%, while the percentage of outliers assuming value 1 increases to about 50%. This property is very interesting. Indeed, while the distribution of the APOA4 attribute is highly unbalanced in the inlier subpopulations, since the majority of the inliers assume value “0” on APOA4, the distribution on the outlier subpopulation is quite uniform.

Another interesting pair returned by EXPREX consists of the property $p =$ TH with explanation $E = \{\text{SIRT3} = 0, \text{mtDNA} = 0, \text{HSP90-}\beta = 1\}$, having exceptionality score $\tau = 0.956$ and involving 30 outliers and 298 inliers. The distributions of the property with empty explanation and with the explanation mined by EXPREX are reported in the following tables:

	D^o	D^i
0	75	665
1	22	210

“TH” without explanation.

	D^o	D^i
0	29	248
1	1	50

“TH” with explanation.

Consider the property TH without explanation. The number of outliers assuming value 0 is 23% and the number of inliers assuming value 0 is 24%. Then, TH is similarly distributed on the inliers and on the outliers and, consequently, the score τ is equal to 0.1848, a rather small value. Conversely, by focusing on the individuals selected by the explanation, the number of inliers slightly decreases to 16%, while the number of outliers drastically decreases to 3%.

It is worth noting that, while for the human analyst it is in general impractical to single out all the relevant property/explanations, due to the enormous size of the search space (consisting of all potential property-explanation pairs), the EXPREX algorithm was able to automatically detect exceptional pairs, and, thus, to reveal knowledge hidden in the data. As an example, the two above discussed pairs have no counterpart in the analysis reported in [4]. The pairs singled out by EXPREX showed a clear significant effect on characterizing longevous individuals, suggesting a direction for a more in-depth analysis to be conducted on the genetic side by biologists.

6 Conclusions

This work aimed at providing a contribution towards the design of automatic methods for the discovery of properties characterizing a small group of outlier individuals as opposed to the whole population of “normal” individuals. In particular, we have introduced the concept of exceptional explanation-property pair and have discussed the significance of the associated knowledge. The innovativeness of the approach has been illustrated by highlighting the substantial differences with related techniques. Moreover, we have defined the concept of exceptionality score, which measures the badness of fit of the values assumed by the outliers with respect to the probability distribution associated with the values assumed by the inliers. Suitable exceptionality scores have been introduced for both numeric and categorical attributes. These scores have been shown, from both the analytical and the empirical point of view, to be effective for small samples, as outlier sets usually are. Thus, our method has been explicitly conceived to deal with rare sub-populations. This fact represents a peculiarity of the method as opposed to related approaches found in the literature. We have designed an algorithm, called EXPREX, which efficiently discovers exceptional explanation-property pairs. Finally, we have shown that our technique is able to provide knowledge characterizing in a natural manner outlier groups, as confirmed by the experimental campaign we have reported in the paper.

REFERENCES

[1] F. Angiulli, F. Fassetti, and L. Palopoli, "Detecting outlying properties of exceptional objects," *ACM Trans. Database Syst.*, vol. 34, no. 1, 2009.

[2] K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arised from random sampling," *Philosophical Magazine Series 5*, vol. 50, pp. 157–175, 1900.

[3] T. W. Anderson, "On the distribution of the two-sample cramér-von mises criterion," *The Annals of Mathematical Statistics (Institute of Mathematical Statistics)*, vol. 33, no. 3, pp. 1148–1159, 1962.

[4] G. Passarino, A. Montesanto, S. Dato, S. Giordano, F. Domma, V. Mari, E. Feraco, and G. D. Benedictis, "Sex and age specificity of susceptibility genes modulating survival at old age," *Human Heredity (Int. Journal of Human and Medical Genetics)*, vol. 62, no. 4, pp. 213–220, 2006.

[5] L. U. Gerdes, B. Jeune, K. A. Ranberg, H. Nybo, and J. W. Vaupel, "Estimation of apolipoprotein e genotype-specific relative mortality risks from the distribution of genotypes in centenarians and middle-aged men: apolipoprotein e gene is a "frailty gene", not a "longevity gene"," *Genetic Epidemiology*, vol. 19, no. 3, pp. 202–210, 2000.

[6] E. Lehmann, *Testing Statistical Hypotheses*, 3rd ed. Springer, 2005.

[7] R. A. Fisher, "On the interpretation of χ^2 from contingency tables, and the calculation of p," *Journal of the Royal Statistical Society*, vol. 85, no. 1, pp. 87–94, 1922.

[8] *Statistical Methods for Research Workers*. Oliver and Boyd, 1954.

[9] S. Stigler, "Fisher and the 5% level," *Chance*, vol. 21, no. 4, 2008.

[10] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artif. Intell. Rev.*, vol. 22, no. 2, pp. 85–126, 2004.

[11] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: special issue on learning from imbalanced data sets," *SIGKDD Explorations*, vol. 6, no. 1, pp. 1–6, 2004.

[12] D. Tax, "One-class classification," Ph.D. dissertation, Delft University of Technology, 2001.

[13] E. Knorr and R. Ng, "Algorithms for mining distance-based outliers in large datasets," in *Procs of VLDB-98*, 1998, pp. 392–403.

[14] F. Angiulli and C. Pizzuti, "Outlier mining in large high-dimensional data sets," *IEEE Transaction on Knowledge and Data Engineering*, vol. 17, no. 2, pp. 203–215, February 2005.

[15] F. Angiulli and F. Fassetti, "Dolphin: an efficient algorithm for mining distance-based outliers in very large datasets," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 3, no. 1, p. Article 4, March 2009.

[16] M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander, "Lof: Identifying density-based local outliers," in *SIGMOD*, 2000, pp. 93–104.

[17] S. Papadimitriou, H. Kitagawa, P. Gibbons, and C. Faloutsos, "Loci: Fast outlier detection using the local correlation integral," in *Procs of ICDE*, 2003, pp. 315–326.

[18] F. Angiulli, G. Greco, and L. Palopoli, "Outlier detection by logic programming," *ACM Trans. Comput. Log.*, vol. 9, no. 1, 2007.

[19] F. Angiulli, R. Ben-Eliyahu-Zohary, and L. Palopoli, "Outlier detection using default reasoning," *Artificial Intelligence (AIJ)*, vol. 172, no. 16–17, pp. 1837–1872, November 2008.

[20] F. Angiulli and F. Fassetti, "Outlier detection using inductive logic programming," in *ICDM*, 2009, pp. 693–698.

[21] G. Dong and J. Li, "Efficient mining of emerging patterns: Discovering trends and differences," in *KDD*, 1999, pp. 43–52.

[22] X. Zhang, G. Dong, and K. Ramamohanarao, "Exploring constraints to efficiently mine emerging patterns from large high-dimensional datasets," in *KDD*, 2000, pp. 310–314.

[23] P. K. Novak, N. Lavrac, and G. I. Webb, "Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining," *Journal of Machine Learning Research*, vol. 10, pp. 377–403, 2009.

[24] J. Li, G. Dong, and K. Ramamohanarao, "Making use of the most expressive jumping emerging patterns for classification," *Knowledge and Information Systems*, vol. 3, no. 2, pp. 1–29, 2001.

[25] J. Bailey, T. Manoukian, and K. Ramamohanarao, "Classification using constrained emerging patterns," in *Advances in Web-Age Information Management*, ser. Lecture Notes in Computer Science. Springer, 2003, vol. 2762, pp. 226–237.

[26] S. D. Bay and M. J. Pazzani, "Detecting change in categorical data: mining contrast sets," in *KDD*, 1999, pp. 302–306.

[27] W. Klösgen, "Explora: A multipattern and multistrategy discovery assistant," in *Advances in Knowledge Discovery and Data Mining (KDD)*, 1996, pp. 249–271.

[28] S. D. Bay and M. J. Pazzani, "Detecting group differences: Mining contrast sets," *Data Mining and Knowledge Discovery*, vol. 5, no. 3, pp. 213–246, 2001.

[29] K. Ramamohanarao, J. Bailey, and H. Fan, "Efficient mining of contrast patterns and their applications to classification," in *ICISIP*, 2005, pp. 39–47.

[30] S. Wrobel, "An algorithm for multi-relational discovery of subgroups," in *PKDD*, 1997, pp. 78–87.

[31] H. Grosskreutz and S. Rüping, "On subgroup discovery in numerical domains," *Data Mining and Knowledge Discovery*, vol. 19, no. 2, pp. 210–226, 2009.

[32] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[33] G. M. Weiss, "Mining with rarity: a unifying framework," *SIGKDD Explorations Newsletter: Special issue on learning from imbalanced datasets*, vol. 6, no. 1, pp. 7–19, 2004.

[34] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *SIGMOD*. New York, NY, USA: ACM, 1993, pp. 207–216.

[35] A. Asuncion and D. Newman, "UCI machine learning repository," 2007. [Online]. Available: http://www.ics.uci.edu/∼mlearn/MLRepository.html