dataminingknowledgediscovery manuscript No.
(will be inserted by the editor)

Model-based Exception Mining for Object-Relational Data

Fatemeh Riahi · Oliver Schulte

Received: date / Accepted: date

Abstract This paper develops model-based exception mining and outlier detection for the case of object-relational data. Object-relational data represent a complex heterogeneous network, which comprises objects of different types, links among these objects, also of different types, and attributes of these links. We follow the well-established exceptional model mining (EMM) framework, which has been previously applied for subgroup discovery in propositional data; our novel contribution is to develop EMM for relational data. EMM leverages machine learning models for exception mining: An object is exceptional to the extent that a model learned for the object data differs from a model learned for the general population. In relational data, EMM can therefore be used for detecting single outlier or exceptional objects. We combine EMM with state-of-the-art statistical-relational model discovery methods for constructing a graphical model (Bayesian network), that compactly represents probabilistic associations in the data. We investigate several outlierness metrics, based on the learned object-relational model, that quantify the extent to which the association pattern of a potential outlier object deviates from that of the whole population. Our method is validated on synthetic data sets and on real-world data sets about soccer and hockey matches, IMDb movies and mutagenic compounds. Compared to baseline methods, the EMM approach achieved the best detection accuracy when combined with a novel outlinerness metric. An empirical evaluation on soccer and movie data shows a strong correlation between our novel outlierness metric and success metrics: Individuals that our metric marks out as unusual tend to have unusual success.

Keywords Outlier Detection · Exception Mining · Statistical-Relational Learning · Bayesian network · Likelihood Ratio · Network Data

This work was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada.

Simon Fraser University
E-mail: sriahi@sfu.ca ; oschulte@cs.sfu.ca

1 Introduction: Exception Mining for Relational Data

Exception mining and outlier detection are important data analysis tasks in many domains. For relational data, exception mining can be applied to support outlier detection, where statistical deviations are viewed as due to a node or entity being genuinely exceptional, rather than due to statistical noise in the data. Statistical approaches to unsupervised outlier detection leverage a generative model of the data Aggarwal (2013). The generative model represents normal behavior. An individual object is deemed an outlier if the model assigns sufficiently low likelihood to generating it. We describe a new method, based on the exceptional model mining (EMM) framework (Duivesteijn et al., 2016), for extending statistical outlier detection with generative models to the case of object-relational data.

Data Type. The object-relational data model is one of the main data models for structured data (Koller and Pfeffer, 1997). The main characteristics of objects that we utilize in this paper are the following.

Object Identity. Each object has a unique identifier that is the same across contexts. For example, a player has a name that identifies him in different matches.

Class Membership. An object is an instance of a class, which is a collection of similar objects. Objects in the same class share a set of attributes. For example, van Persie is a player object that belongs to the class striker, which is a subclass of the class player.

Object Relationships. Objects are linked to other objects. Both objects and their links have attributes.

A common type of object relationship is a component relationship between a complex object and its parts. For example, a match links two teams, and each team fields a set of players for that match. A difference between relational and vectorial data is therefore that an individual object is characterized not only by a list of attributes, but also by its links and by attributes of the objects linked to it. We refer to the substructure comprising this information as the *object data*. Equivalent terms are “egonet” from network analysis (Akoglu et al., 2015) and “interpretation” (Maervoet et al., 2012). The *relational outlier detection problem is to identify objects whose object data differ from the general population/class.* Our approach to this problem combines EMM with statistical-relational model discovery (Getoor and Taskar, 2007).

Approach: Exceptional Model Mining + Statistical-Relational Learning. In the EMM framework, a subgroup is exceptional to the extent that a model learned from data for the subgroup deviates from a model learned for the general population. Figure 1 illustrates an EMM blueprint. EMM therefore leverages the extensive work on model learning for exception mining. Relational EMM leverages the extensive work on statistical-relational model learning for relational exception mining. In the propositional setting with i.i.d. data (independent and identically distributed), each object is represented by a single data row, and it is meaningless to learn a model for a single object. Instead, EMM is applied to identify exceptional subgroups of objects. With relational data, each object is represented by its own data set, and it is meaningful to apply EMM to identify single exceptional objects—a subgroup of size 1. Therefore *relational EMM can be utilized for relational outlier detection*; this paper evaluates relational EMM for this task.

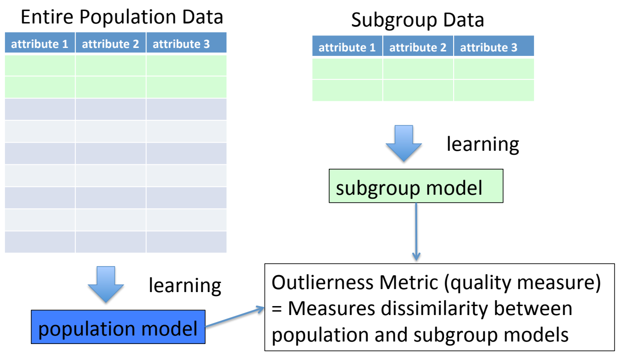


Fig. 1 A general schema for exceptional model mining for propositional i.i.d. data

Model Type. EMM is an inclusive framework in that any model type of interest can be utilized. In this paper we focus on first-order Bayesian networks (BNs) (Wang et al., 2008; Poole, 2003; Kimmig et al., 2014), but our definitions apply to other log-linear models such as Markov logic networks (Domingos and Richardson, 2007). A class-model Bayesian network (BN) structure is learned with data for the entire population. The nodes in the BN represent attributes for links, of multiple types, and attributes of objects, also of multiple types. To learn the BN model, we apply techniques from statistical-relational learning, a recent field that applies and extends AI and machine learning to relational data (Getoor and Taskar, 2007; Schulte and Khosravi, 2012; Domingos and Lowd, 2009). For a learned BN structure, we compare two parameter vectors (see Figure 6 below): The *class model parameters* are estimated from the entire population data, and the *object model parameters* are estimated from the object data.

Quality Measure = Outlierness Metric. For a given model type, quantifying the extent to which an object model deviates from the population class model is the main research question in EMM (Duivesteijn et al., 2016). A computational method for measuring this extent is called *a quality measure*; when we apply EMM to relational outlier detection, we also refer to it as an *outlierness metric*. We propose utilizing the relational log-linear likelihood function (Kimmig et al., 2014; Schulte, 2011) that measures how well a parametrized statistical-relational model fits an input database. The *likelihood ratio* is the difference in log-likelihood between the class model and the object model, when applied to the object data. Assuming maximum likelihood estimates, it is equivalent to the Kullback-Leibler divergence (KLD) between class and object models. While our evaluation indicates that KLD is a good outlierness metric, we propose improving it further by applying two transformations: (1) a mutual information decomposition, and (2) replacing log-likelihood differences by log-likelihood distances. We refer to the resulting novel score as the *expected log-likelihood distance* (ELD). A Taylor series analysis shows that this score is approximated by the L1-norm (total variation distance) between the object and class data distributions.

Evaluation. Our code and data sets are available on-line at (Riahi and Schulte, 2015b). We evaluate relational EMM outlierness metrics in three ways.

Detection Accuracy. Our performance evaluation follows the design of previous outlier detection studies (Gao et al., 2010; Aggarwal, 2013), where the methods are scored against a test set of known outliers. We use three synthetic and four real-world data sets, from the UK Premier Soccer League, the Internet Movie Database (IMDb), the National Hockey League, and Mutagenesis. On the synthetic data we have known ground truth. For the real-world data sets, we use a one-class design, where one object class is designated as normal and objects from outside the class are the outliers. For example, we compare goalies as outliers against the class of strikers as normal objects. Compared to six baseline methods, the EMM-based scores (likelihood ratio and expected log-likelihood distance) achieve the top detection accuracy on all three synthetic data sets, and on 4 out of 5 real-world data sets.

Case Studies. We also offer case studies where we assess whether individuals that our score ranks as highly unusual in their class are indeed unusual. The case studies illustrate that the EMM scores are *easy to interpret*, because the Bayesian network provides a local decomposition of log-likelihood differences. Interpretability is very important for users of an outlier detection method as there is often no ground truth to evaluate outliers.

Correlation with Success. We compare the likelihood-based outlierness metrics to independent success metrics for a given domain. Success rankings are one of the most interesting features to users. Our reasoning is that high success is an independent metric that indicates an unusual individual. So a correlation between an outlierness metric and success is an independent validation of the meric, and also shows that it points to meaningful and interesting outliers.

Contributions Our main contributions may be summarized as follows.

1. We develop the first application of the EMM framework to outlier detection for structured relational data based on a probabilistic model.
2. We define a new model-based outlierness score, derived from a novel model likelihood comparison, the expected log-likelihood distance.

Paper Organization We review related work on outlier detection for structured data, then background about Bayesian networks for relational data. Then we introduce likelihood-ratio based outlierness scores for relational EMM, including our novel log-likelihood distance. Empirical evaluation compares EMM and aggregation-based approaches to relational outlier detection, with respect to three synthetic and four real-world data sets.

2 Related Work

Outlier detection is a densely researched field; for surveys please see Aggarwal (2013) and Akoglu et al. (2015). Figure 2 provides a tree picture of where our method is situated with respect to other outlier detection methods and other data models. Our method falls in the category of *unsupervised* statistical model-based approaches. To our knowledge, ours is the first model-based method tailored for object-relational data. Like EMM and other model-based approaches, it detects *global outliers.* Aggarwal (2013) defines a global outlier to be a data point that notably deviates from the rest of the population. We review relevant approaches from two different data models, the most common atomic object model—where data is represented by vectors—and structured data models. A conference presentation of part of our work is available (Riahi and Schulte, 2015a).

a) Attribute Vector Data Model: By far most work on outlier detection considers atomic objects with flat feature vectors. This leads to an impedance mismatch: The required input format for these outlier detection methods is a single data matrix, not a structured data set. For example, one cannot provide a relational database as input. This mismatch is not simply a question of choosing a file format, but instead reflects a different underlying data model: complex objects with both attributes and component objects vs. atomic objects with attributes only. It is possible to "flatten" structured data by converting it to unstructured feature vectors, for instance by using aggregate functions. We evaluated the aggregation approach in this paper by applying three standard methods for outlier detection.

Work on atomic contextual outliers (Tang et al., 2013) is like ours in that it considers the distinctness of a target individual from a reference class. A reference class is not specified for each object, but is constructed as part of outlier detection. Our work could be combined with a class discovery approach by providing a score of how informative the inferred classes are.

b) Structured Data Models: We discuss related techniques in three types of structured data models: SQL (relational), XML (hierarchical), and OLAP (multi-dimensional).

For relational data, many outlier detection approaches aim to discover rules that represent the presence of anomalous associations for an individual or the absence of normal associations (Maervoet et al., 2012; Gao et al., 2010). In these approaches what we call the object data is usually referred to as an interpretation, so the problem is to score outlier interpretations (Maervoet et al., 2012). The survey by Novak et al. (2009) unifies within a general rule search

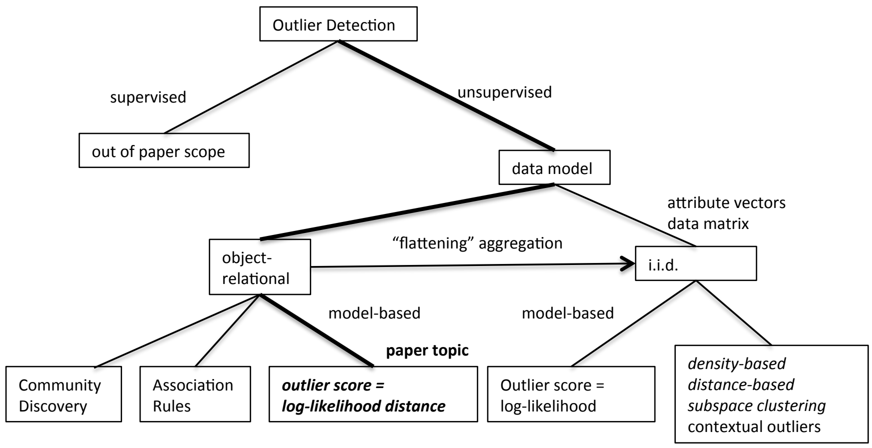


Fig. 2 A tree structure for related work on outlier detection for structured data. A path specifies an outlier detection problem, the leaves list major approaches to the problem. Approaches in italics appear in experiments.

framework three important related tasks: i) exception mining, which looks for associations that characterize unusual cases, ii) subgroup mining, which looks for associations characterizing important subgroups, and iii) contrast space mining, which looks for differences between classes. Another rule-based approach uses Inductive Logic Programming techniques (Angiulli et al., 2004). While local rules are informative, they are not based on a global statistical model and do not provide a single outlierness score for each individual. As we show in our case studies (Section 8.3), the conditional probability parameters in a Bayesian network can be used to extract rules from the BN model (of the form $parent\_values \rightarrow child\_value$). The distance-based algorithms designed for outlier detection can be applied to the relational case. Horváth et al. (2001) introduced a similarity measure for the first-order instance-based learner RIBL that uses the edit distance to compute the distance between lists and terms. This distance has been used for the clustering task by Kirsten et al. (2001). To the best of our knowledge, distance-based relational learning has not been used for outlier detection. One-class classification can be viewed as a special case of outlier detection. Khot et al. (2014) introduced a non-parametric relational one-class classification based on first-order trees. They proposed a tree-based distance metric to discover new relational features and to differentiate relational examples. Khot *et al.* emphasize that their method is not appropriate for outlier detection because outlier detection makes different assumptions about the data than they do (e.g., outliers are isolated and far from other points).

Propositionalization summarizes the multi-relational data into a single data table and can be used for outlier detection and classification tasks (Kramer et al., 2000; Perovsek et al., 2013; Kuvzelka and Zelezny, 2008; Riahi and Schulte, 2016; Anderson and Pfahringer, 2008).

A latent variable approach in information networks ranks potential outliers in reference to the latent communities inferred by network analysis (Gao et al., 2010). Our model also aggregates information from entities and links of different types, but does not assume that different communities have been identified.

Koh et al. (2008) propose a method for hierarchical structures represented in XML document trees. Their aim is to identify feature outliers, not class outliers as in our work. Also, they use aggregate functions to convert the object hierarchy into feature vectors. Their outlierness score is based on local correlations, and they do not construct a statistical model.

The multi-dimensional data model defines numeric measures for a set of dimensions. The differences in the two data models mean that multi-dimensional outlier detection models (Sarawagi et al., 1998) do not carry over to object-relational outlier detection. (1) The object data model allows but does not require any numeric measures. In our data sets, all features are discrete. Nor do we assume that it is possible to aggregate numeric measures to summarize lower-level data at higher levels. (2) In scoring a potential outlier object, our method considers other objects *both* below and above the target object in the component hierarchy. OLAP exploration methods consider only cells below or at the same level as the target cell. For example, in scoring a player, our method would consider features of the player's team. Also, the *ELD* outlierness score of an object is not determined by the outlierness scores of its components, in contrast to approaches derived from the work of Sarawagi et al. (1998). They use values such as the most unusual cell that is below a target cell. (3) Our approach models a joint distribution over features, exploiting correlations among features. Most of the OLAP-based methods consider only a single numeric measure at a time, not a joint multi-variate model.

3 Background: Bayesian Networks for Relational Data

We adopt the first-order Bayesian network (PBN) formalism (Poole, 2003) that combines Bayesian networks with logical syntax for expressing relational concepts.

3.1 Relational Data

PBNs use a term-based notation for combining logical and statistical concepts (Poole, 2003; Kimmig et al., 2014; Wang et al., 2008). Table 1 summarizes our notation. A functor is a function or predicate symbol. Each functor has a set of values (constants) called the **domain** of the functor. The range of a **predicate** is $\{T, F\}$. Predicates are usually written with uppercase Roman letters, other terms with lowercase letters. A predicate of arity at least two is a **relationship** functor. Relationship functors specify which objects are linked. Other functors represent **features** or **attributes** of an object or a tuple of

Symbol	Definition
$a, b, a_1, b_1, \ldots$	Constant
$A, B, T, M, \ldots$	First-order variable
$f, g, \ldots,$	Functor, function symbol
$f(A, \ldots, A_k)$	First-order term
$f(A = a_1, \ldots, A_k = a_k)$	Ground term
$\mathcal{D}$	Relational database
$\mathcal{D}_C$	Database for the entire class of objects
$\mathcal{D}_o$	Restriction of the input database to the target object
$V(A, ..., B)$	A first-order random variable
$\boldsymbol{V}$	A set of first-order random variable
$\boldsymbol{V} = \mathbf{v}$	Joint assignment of values to a set of FORVs
$P(\boldsymbol{V} = \mathbf{v}) \equiv P(\mathbf{v})$	Joint probability that each variable V_i takes on value $\mathbf{v}_i$
$\#_{\mathcal{D}}(\boldsymbol{V} = \mathbf{v})$	Count of groundings that satisfy the assignment
$A \backslash v$	Ground a first-order variable
B	A Bayesian network structure
B_C	A Bayesian network structure learned with $\mathcal{D}_C$ as the input database
$\boldsymbol{\theta}_C$	Parameters learned for B_C using $\mathcal{D}_c$ as the input database
$\boldsymbol{\theta}_o$	Parameters learned for B_C using $\mathcal{D}_o$ as the input database
$\mathbf{pa}_i$	Parent of node i

Table 1 Notation and Definition

objects (i.e., of a relationship). A **population** is a set of objects. A **term** is of the form $f(\sigma_1, \ldots, \sigma_k)$ where f is a functor and each σ_i is a first-order variable or a constant denoting an object. A term is **ground** if it contains no first-order variables; otherwise it is a first-order term. In the context of a statistical model, we refer to first-order terms as **first-order random variables** (FORVs) (Wang et al., 2008; Schulte and Gholami, 2017). A **grounding** replaces each first-order variable in a term by a constant; the result is a ground term. A grounding may be applied simultaneously to a set of terms. A relational database $\mathcal{D}$ specifies the values of all ground terms.

Consider a joint assignment $P(\boldsymbol{V} = \mathbf{v})$ of values to a set of FORVs $\boldsymbol{V}$. The *grounding space* of the FORVs is the set of all possible grounding substitutions, each applied to all FORVs in $\boldsymbol{V}$. The *count* of groundings that satisfy the assignment with respect to a database $\mathcal{D}$ is denoted by $\#_{\mathcal{D}}(\boldsymbol{V} = \mathbf{v})$. The **database frequency** $P_{\mathcal{D}}(\boldsymbol{V} = \mathbf{v})$ is the grounding count divided by the number of all possible groundings (Halpern, 1990; Schulte et al., 2014).

Example. The Opta data set represents information about Premier League data (Sec. 7.2). The basic populations are teams, players, matches, with corresponding first-order variables T, P, M. As shown in Table 2, the groundings count can be visualized in terms of a groundings table (Schulte et al., 2014), also called a universal schema (Riedel et al., 2013). The first three column headers show first-order variables ranging over different populations. The remaining columns represent terms. Each row represents a single grounding and the values of the ground terms defined by the grounding. In terms of the grounding table, the grounding count of a joint assignment is the number of rows that satisfy the conditions in the joint assignment. The database frequency is the grounding count divided by the total number of rows in the groundings table.

Table 4 shows an example of computing frequencies. Counts are based on the 2011-2012 Premier League Season. We count only groundings $(team, match)$ such that *team* plays in *match*. Each team, including Wigan Athletic, appears in 38 matches. The total number of team-match pairs is $38 \times 20 = 760$.

MatchId M	TeamId T	PlayerId P	TimePlayed(P,M)	ShotEff(T,M)	result(T,M)
117	WA	McCarthy	90	0.53	win
148	WA	McCarthy	85	0.57	loss
15	MC	Silva	90	0.59	win
...	...	...	...	...	...

Table 2 Sample Population Data Table (Soccer).

MatchId M	TeamId $T = WA$	PlayerId P	TimePlayed(P,M)	ShotEff(WA,M)	result(WA,M)
117	WA	McCarthy	90	0.53	win
148	WA	McCarthy	85	0.57	loss
...	WA	...	...	...	...

Table 3 Sample Object Data Table, for team $T = WA$; all rows have WA in the T column.

A novel aspect of our paper is that we *learn model parameters for specific objects as well as for the entire population.* The appropriate **object data table** is formed from the population data table by restricting the relevant first-order variable to the target object. For example, the object database for target Team *WiganAthletic*, forms a subtable of the data table of Table 2 that contains only rows where TeamID = WA; see Table 3. In database terminology, an object database is like a view centered on the object.

3.2 Bayesian Networks

A Bayesian network (BN) structure B is a Directed Acyclic Graph (DAG) whose nodes comprise a set of random variables (Pearl, 1988). Depending on context, we interchangeably refer to the nodes and variables of a BN, collectively denoted by $\boldsymbol{V} = \{V_1, \ldots, V_n\}$. The possible values of V_i are enumerated as $\{v_{i1}, \ldots, v_{ir_i}\}$. The notation $P(V_i = v) \equiv P(v)$ denotes the probability of variable V_i taking on value v. We also use the vector notation $P(\boldsymbol{V} = \mathbf{v}) \equiv P(\mathbf{v})$ to denote the joint probability that each variable V_i takes on value $\mathbf{v}_i$.

Database	Count or $\#_D(\boldsymbol{V} = \mathbf{v})$	Frequency or $P_D(\boldsymbol{V} = \mathbf{v})$
Population	76	$76/760 = 0.10$
Wigan Athletic	7	$7/38 = 0.18$

Table 4 Example of Grounding Count and Frequency in Premier League Data, for the conjunction $passEff(T, M) = hi, shotEff(T, M) = hi, Result(T, M) = win$.

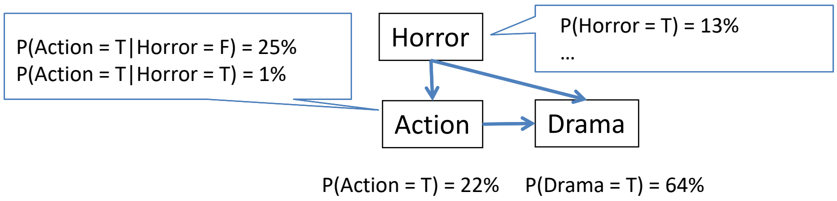


Fig. 3 Example of joint and marginal probabilities computed from a simple Bayesian network structure that was learned from the *Movies* table in the IMDb data set described in Section 7.2. The parameters were estimated from the IMDb data set. The conditional probability parameters for the *Drama* node are not shown. The Bayesian network shows that movie genres are largely but not completely exclusive. For instance, among horror movies, only 1% are also classified as action movies. The marginal probabilities are the base rate frequencies of horror, action, and drama movies, which are respectively 13%, 22%, 64%.

The conditional probability parameters of a Bayesian network specify the distribution of a child node given an assignment of values to its parent nodes. For an assignment of values to its nodes, a BN defines the joint probability as the product of the conditional probability of the child node value given its parent values, for each node in the network. This means that the log-joint probability can be *decomposed* as the node-wise sum

$$\ln P(\boldsymbol{V} = \mathbf{v}; B, \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \theta(\mathbf{v}_i | \mathbf{v}_{\mathbf{pa}_i}) \tag{1}$$

where $\mathbf{v}_i$ is the assignment of values to node V_i, and $\mathbf{v}_{\mathbf{pa}_i}$ is the assignment of values to the parents of V_i, as determined by the assignment $\mathbf{v}$. To avoid difficulties with $\ln(0)$, here and below we assume that joint distributions are positive everywhere. Since the parameter values $\boldsymbol{\theta}$ for a Bayesian network define a joint distribution over its nodes, they therefore entail a marginal, or unconditional, probability for a single node. We denote the **marginal probability** that node V has value v as $P(V = v; B, \boldsymbol{\theta}) \equiv \theta(v)$. In the following we use the term Bayesian network model to refer to a network structure with parameters (i.e., a pair $(B, \boldsymbol{\theta})$); for brevity, we also use the terms “Bayesian network” or “model”.

Example. Figure 3 shows an example of a Bayesian network model and associated conditional and marginal probabilities.

3.3 Target Model: Bayesian Networks for Relational Data

A **first-order Bayesian Network Structure** (FOBN) is a Bayesian network structure whose nodes are FORVs (Poole, 2003; Wang et al., 2008; Schulte and Gholami, 2017). The relationships and features in an object database define a set of nodes for a first-order Bayesian network. I.i.d. data represented in a

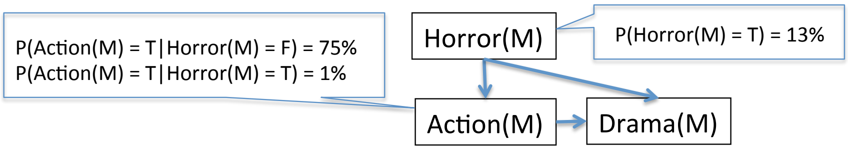


Fig. 4 The Bayesian network from Figure 3, where the node names are expanded using the syntax of first-order random variables.

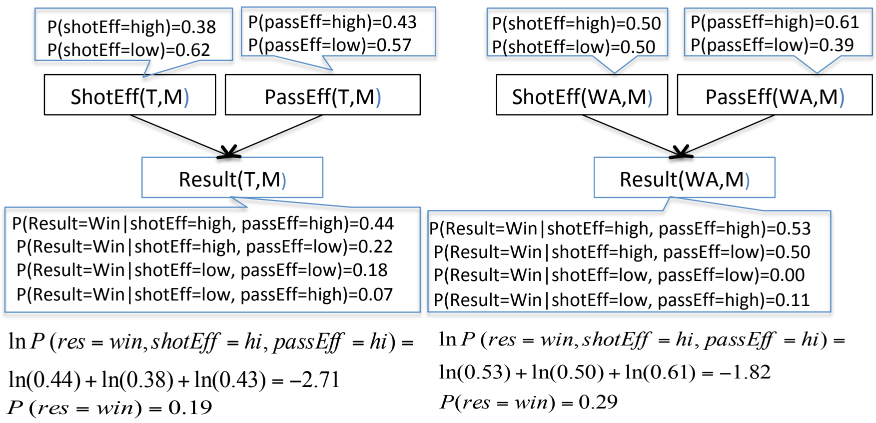


Fig. 5 Example of joint and marginal probabilities computed from a toy Bayesian network structure. The parameters were estimated from the Premier League data set. (left): A class model Bayesian network B_c for all teams with class parameters $\boldsymbol{\theta}_c$. The first-order variable T ranges over teams, and the first-order variable M over matches. For compactness, conditional probabilities omit the first-order variables and constants. (right): The same Bayesian network structure with object parameters $\boldsymbol{\theta}_o$ learned for Wigan Athletic ($T = WA$).

single table can be viewed as a *special limiting case of multi-relational data with no relationships* (Nickel et al., 2016; Knobbe, 2006). Syntactically, this means that columns in an i.i.d. data table represent unary functors, where the relevant population is assumed to be clear from the context rather than explicitly specified as a first-order variable. Figure 4 illustrates how the usual syntax for i.i.d. Bayesian networks is a special case of the FOBN syntax.

Figure 5 shows a first-order Bayesian network for the Premier League domain that is truly relational in that the functors depend on more than one population variable. For example, shot efficiency does not depend on a match only, but also depends on specifying a team. The BN product formula (1) can be applied to any FOBN to compute (estimated) frequencies. In the case of a truly relational FOBN, as in Figure 5, the FOBN can be viewed as representing database frequencies (rather than data table frequencies as in Figure 4). Using Getoor's terminology, the FOBN can be viewed as a Statistical-Relational Model (SRM) (Getoor, 2001; Schulte et al., 2014; Schulte and Gholami, 2017).

3.4 Likelihood Score for first-order Bayesian Networks.

A standard method for applying a generative model assumes that the generative model represents normal behavior since it was learned from the entire population. An object is deemed an outlier if the model assigns sufficiently low likelihood to generating its features (Cansado and Soto, 2008). This likelihood method is an important baseline for our investigation. The other outlierness scores we consider are improved variants of the likelihood approach. Defining a likelihood for relational data is more complicated than for i.i.d. data, because an object is characterized not only by a feature vector, but by an object database. We employ the previously defined normalized relational log-likelihood score (Schulte, 2011; Xiang and Neville, 2011), which can be computed as follows for a given Bayesian network and database.

$$LOG(\mathcal{D}, B, \boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{j=1}^{r_i} \sum_{\mathbf{pa}_i} P_{\mathcal{D}}(v_{ij}, \mathbf{pa}_i) \ln \theta(v_{ij}|\mathbf{pa}_i) \tag{2}$$

Equation (2) has the same form of the standard BN log-likelihood function for vectorial i.i.d. data (de Campos, 2006), except that parent-child instantiation counts are standardized to be proportions (Schulte, 2011; Schulte and Gholami, 2017). For the notation please refer to Table 1. In general the exponential of the score (2) is a pseudo-likelihood, in that it need not sum to 1 over all relational databases $\mathcal{D}$.[1] The equation can be read as follows.

1. For each parent-child configuration, use the conditional probability of the child given the parent.
2. Multiply the logarithm of the conditional probability by the database frequency of the parent-child configuration.
3. Sum this product over all parent-child configurations and all nodes.

The maximum of the relational log-likelihood (2) is given by the empirical database frequencies (Schulte, 2011, Prop.3.1.). In all our experiments we use these maximum likelihood parameter estimates.

Example. Consider the class-level BN of Figure 5(left). The family configuration

$$passEff(T, M) = high, shotEff(T, M) = high, Result(T, M) = win$$

contributes one term to the log-likelihood score For the population database $\mathcal{D}$, this term is

$$P_D(passEff(T, M) = high, shotEff(T, M) = high, Result(T, M) = win) \times$$
$$\ln P(Result(T, M) = win|passEff(T, M) = high, shotEff(T, M) = high) =$$
$$0.1 \times \ln(0.44) = -0.08$$

[1] The conditions under which the score of Equation (2) sums to 1 over all relational data sets are discussed by Schulte (2011). Briefly, if the first-order Bayesian Network is viewed as a template for an unrolled ground network, then (1) the ground network cannot contain cycles, and (2) each ground network cannot have multiple parent instantiations; see also (Heckerman et al., 2007).

where the database frequency 0.1 is computed as in Table 4 and the conditional probability 0.44 is the parameter specified by the Bayesian network of Figure 5.

For the Wigan Athletic data $\mathcal{D}_{WA}$, the corresponding term is

$$P_{\mathcal{D}_{WA}}(\mathit{passEff}(\mathit{WA}, M) = \mathit{high}, \mathit{shotEff}(\mathit{WA}, M) = \mathit{high}, \mathit{Result}(\mathit{WA}, M) = \mathit{win}) \times \ln P(\mathit{Result}(WA, M) = \mathit{win} | \mathit{passEff}(\mathit{WA}, M) = \mathit{high}, \mathit{shotEff}(\mathit{WA}, M) = \mathit{high}) = 0.18 \times \ln(0.44) = -0.14$$

where the database frequency 0.18 is computed as in Table 4 and the conditional probability 0.44 is the parameter specified by the Bayesian network of Figure 5(left).

4 Exceptional Model Mining for Relational Data

This section describes our approach to applying the EMM framework to relational data, using the following notation.

- $\mathcal{D}_C$ is the database for the entire class of objects; cf. Table 2. This database defines the **class distribution** $P_C \equiv P_{\mathcal{D}_c}$.
- $\mathcal{D}_o$ is the restriction of the input database to the target object; cf. Table 3. This database defines the **object distribution** $P_o \equiv P_{\mathcal{D}_o}$.
- B_C is a Bayesian network structure learned with $\mathcal{D}_C$ as the input database. Note that we use the entire population data to learn the Bayesian network structure. Therefore, the Bayesian network structure is the same across different individuals.
- $\boldsymbol{\theta}_C$ resp. $\boldsymbol{\theta}_o$ are parameters learned for B_C using $\mathcal{D}_c$ resp. $\mathcal{D}_o$ as the input database.

Figure 6 illustrates these concepts and the system flow for computing an outlierness score. First, we learn a Bayesian network structure B_C for the entire population using a previous learning algorithm (see Section 7.4). We then compute an *outlierness metric* that evaluates how much the object model deviates from the class model. For a given model type, quantifying the extent to which an object model deviates from the population class model is the main research question in EMM (Duivesteijn et al., 2016). In this paper we examine several outlierness metrics that are based on the likelihood ratio between the object model and the class model. A novel metric that we propose is the **expected log-likelihood distance** (*ELD*). It is defined as follows for each feature i; the total score is the sum of feature-wise scores. Section 5.2 provides example computations.

$$ELD_i = \sum_{j=1}^{r_i} P_o(v_{ij}) \left| \ln \frac{\theta_o(v_{ij})}{\theta_C(v_{ij})} \right| + \tag{3}$$

$$\sum_{j=1}^{r_i} \sum_{\mathbf{pa}_i} P_o(v_{ij}, \mathbf{pa}_i) \left| \ln \frac{\theta_o(v_{ij}|\mathbf{pa}_i)}{\theta_o(v_{ij})} - \ln \frac{\theta_C(v_{ij}|\mathbf{pa}_i)}{\theta_C(v_{ij})} \right|. \tag{4}$$

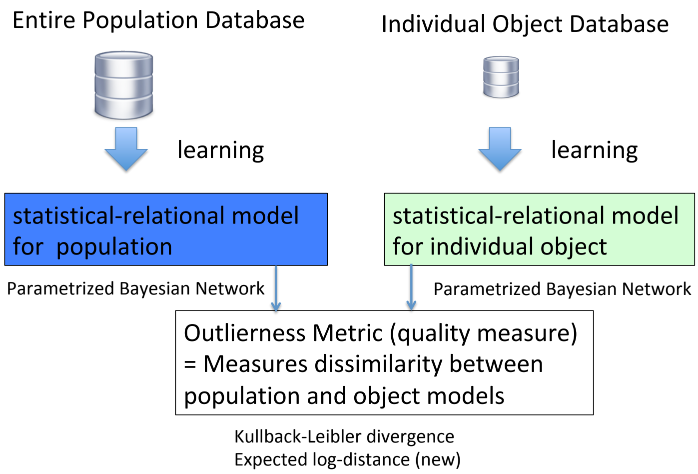


Fig. 6 Exceptional model mining for statistical-relational models (cf. Figure 1). The model class we utilize in this paper are first-order Bayesian networks, with a log-linear likelihood function. As outlierness metrics we consider the standard Kullback-Leibler divergence, and the novel divergence *ELD* described in the text.

4.1 Interpretation

The first sum (3) is the **single-feature** component, where each feature is considered independently of all others. It computes the expected log-distance with respect to the single feature value probabilities between the object and the class models. The second *ELD* sum (4) is the **mutual information component**, based on the mutual information among all features. It computes the expected log-distance with respect to the mutual information of feature value assignments between the object and the class models. Intuitively, the first sum measures how the models differ if we treat each feature in isolation. The second sum measures how the models differ in terms of how strongly parent and child features are associated with each other.

4.2 Motivation

The motivation for the mutual information decomposition is two-fold. (1) *Interpretability.* This is very important for outlier detection. The single-feature components are easy to interpret since they involve no feature interactions. Each parent-child local factor is based on the average relevance of parent values for predicting the value of the child node, where relevance is measured by

$$\ln \frac{\theta(v_{ij}|\mathbf{pa}_i)}{\theta(v_{ij})}.$$

This relevance term is basically the same as the widely used lift measure (Tuffery, 2011), hence is an intuitively meaningful quantity. The *ELD* score compares how relevant a given parent condition is in the object data with how relevant it is in the general class.

(2) *Avoiding cancellations.* For different child-parent configurations, the different components of the *ELD* sum may have different signs. This leads to a partial cancelling of differences between the class and object distributions. Since our goal is to assess the distinctness of an object, *we do not want differences to cancel out.* Taking distances as in Equations (3) and (4) avoids the undesirable loss of information. The next section provides comparison scores and example computations that illustrate the cancelling phenomenon that occurs without adding absolute values. The subsequent section 6 provides a Taylor series analysis for a theoretical understanding of the cancelling phenomenon: We show that without the absolute values (as with *LR*), the first-order term in the Taylor series approximation vanishes, whereas with the absolute values (as with *ELD*), the first-order term is equivalent to the total variation distance (L1 norm).

5 Comparison Scores

We introduce several alternative likelihood-ratio based outlierness scores, following a lesion design where different scores omit different components of our main *ELD* proposal. To illustrate their essential difference with *ELD*, we give toy examples before we evaluate them on full data sets.

5.1 Definition of Comparison outlierness scores

Log-likelihood Ratio Score. Our first comparison score omits the absolute values from the *ELD* score:

$$LR_i = \sum_{j=1}^{r_i} P_o(v_{ij}) \ln \frac{\theta_o(v_{ij})}{\theta_C(v_{ij})} + \sum_{j=1}^{r_i} \sum_{\mathbf{pa}_i} P_o(v_{ij}, \mathbf{pa}_i)(\ln \frac{\theta_o(v_{ij}|\mathbf{pa}_i)}{\theta_o(v_{ij})} - \ln \frac{\theta_C(v_{ij}|\mathbf{pa}_i)}{\theta_C(v_{ij})}).$$

By using the **mutual information decomposition:**

$$\ln \frac{\theta_o(v_{ij}|\mathbf{pa}_i)}{\theta_C(v_{ij}|\mathbf{pa}_i)} = \ln \frac{\theta_o(v_{ij}|\mathbf{pa}_i)}{\theta_o(v_{ij})} - \ln \frac{\theta_C(v_{ij}|\mathbf{pa}_i)}{\theta_C(v_{ij})} + \ln \frac{\theta_o(v_{ij})}{\theta_C(v_{ij})}, \quad (5)$$

it can be shown that the *ELD* score without the absolute values is equivalent to the likelihood ratio, or **log-likelihood difference**:

$$LR(\mathcal{D}_o, B_C, \boldsymbol{\theta}_o) \equiv LOG(\mathcal{D}_o, B_C, \boldsymbol{\theta}_o) - LOG(\mathcal{D}_o, B_C, \boldsymbol{\theta}_C). \quad (6)$$

Assuming maximum likelihood parameter estimation, *LR* is equivalent to the Kullback-Leibler divergence between the class-level and object-level parameters (de Campos, 2006). The log-likelihood difference compares how well the class-level parameters fit the object data with respect to a particular object, vs. how well the object parameters fit the object data. It measures how much the log-conditional probabilities in the class distribution differ from those in the object distribution.

The Feature Divergence Score. The feature divergence outlierness score *FD* uses only part (3) of the *ELD* score. That is, it considers each feature independent of all others. *FD* computes the expected log-distance with respect to the single feature value probabilities between the object and the class models:

$$FD_i = \sum_{i=1}^{n} \sum_{j=1}^{r_i} P_o(v_{ij}) \left| \ln \frac{\theta_o(v_{ij})}{\theta_C(v_{ij})} \right| \tag{7}$$

Log-Likelihood Score. In previous work on applying Bayesian networks to outlier detection for i.i.d. non-relational data, the outlier metric used was the log-likelihood of a data point (Cansado and Soto, 2008). This means that an object was deemed an outlier if the model assigns sufficiently low likelihood to generating its features. For relational data, we use the relational log-likelihood (2) of the target *database*:

$$LOG(\mathcal{D}_o, B_C, \boldsymbol{\theta}_C) = \sum_{i=1}^{n} \sum_{j=1}^{r_i} \sum_{\mathbf{pa}_i} P_{\mathcal{D}}(v_{ij}, \mathbf{pa}_i) \ln \theta(v_{ij}|\mathbf{pa}_i). \tag{8}$$

5.2 Score Computation Examples

We provide three simple examples with only two variables/features that illustrate the computation of the outlierness scores. They are designed so that outliers and normal objects are easy to distinguish, and so that it is easy to trace the behavior of an outlierness score. The examples therefore serve as thought experiments that bring out some key strengths and weaknesses of the model-based outlierness scores we evaluate. Figure 7 describes the BN representation of the examples. For intuition, we can think of a soccer setting, where each match assigns a value to each attribute F_1 and F_2 for each player.

Table 5 illustrates the computation of the scores. Scores for the F_2 feature are computed conditional on $F_1 = 1$. Expectation terms are computed first for $F_2 = 1$, then $F_2 = 0$. The table shows the cancelling effects in *LR*. In the high correlation scenario 7(a), the outlier object has a lower probability than the normal class distribution of *Match_Result = 0* given that *Shot_Efficiency = 1*. Specifically, 0.5 vs. 0.9. The outlier object exhibits a higher probability *Match_Result = 1* than the normal class distribution, conditional on *Shot_Efficiency = 1*; specifically, 0.5 vs. 0.1. In line 1, column 2 of Table 5 the log-ratios ln(0.5/0.9) and ln(0.5/0.1) therefore have different signs. In the low correlation scenario 7(b), the cancelling occurs in the same way, but with the normal and outlier probabilities reversed. The cancelling effect is even stronger for attributes with more than two possible values.

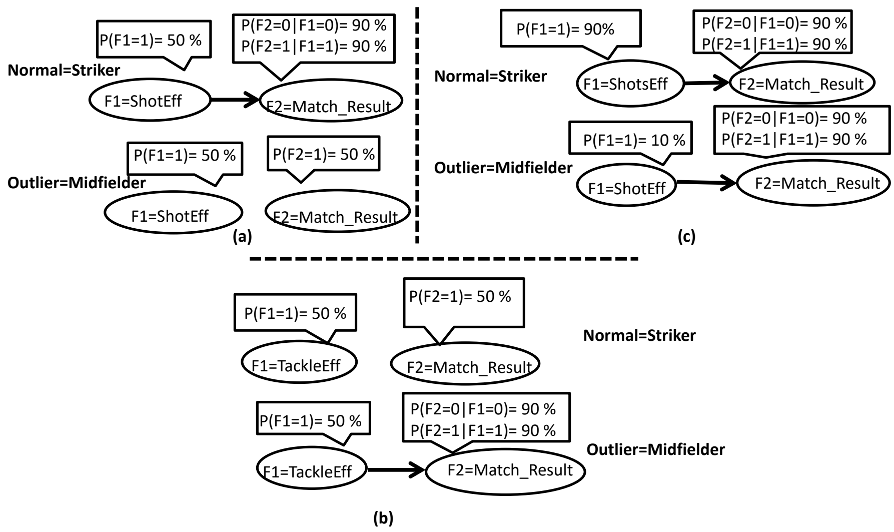


Fig. 7 Illustrative Bayesian networks with two nodes. The networks are not learned from data, but hand-constructed to be plausible for the soccer domain. (a) *High Correlation:* Normal individuals exhibit a strong association between their features, outliers no association. Both normals and outliers have a close to uniform distribution over single features. (b) *Low Correlation:* Normal individuals exhibit no association between their features, outliers have a strong association. Both normals and outliers have a close to uniform distribution over single features. (c) *Single Attributes:* Both normal and outlier individuals exhibit a strong association between their features. In normals, 90% of the time, feature 1 has value 1. For outliers, feature 1 has value 1 only 10% of the time.

Table 5 Example computation of different outlierness scores for outliers given the distributions of Figure 7(a),(b). Our new *ELD* metric contains no negative terms due to the use of absolute values.

Score	$F1 = 1$ Computation	$F2\|F1 = 1$ Computation	Result
LR	$1/2\ln(0.5/0.5) = 0$	$1/4\ln(0.5/0.9) + 1/4\ln(0.5/0.1)$	0.36
FD	$\|\ln(0.5/0.5)\| = 0$	$1/2\|\ln(0.5/0.5)\| + 1/2\|\ln(0.5/0.5)\|$	0
ELD	0 (no parents)	$1/2\|\ln(0.5/0.5)\| + 1/2\|\ln(0.5/0.5)\| + 1/4\|\ln(0.5/0.5) - \ln(0.9/0.5)\| + 1/4\|\ln(0.5/0.5) - \ln(0.1/0.5)\|$	0.79

Table 5(a): High Correlation Case, following Figure 7(a).

Score	$F1 = 1$ Computation	$F2\|F1 = 1$ Computation	Result
LR	$1/2\ln(0.5/0.5) = 0$	$0.5 \cdot 0.9\ln(0.9/0.5) + 0.5 \cdot 0.1\ln(0.1/0.5)$	0.26
FD	$\|\ln(0.5/0.5)\| = 0$	$1/2\|\ln(0.5/0.5)\| + 1/2\|\ln(0.5/0.5)\|$	0
ELD	0 (no parents)	$1/2\|\ln(0.5/0.5)\| + 1/2\|\ln(0.5/0.5)\| + 0.5 \cdot 0.9\|\ln(0.9/0.5) - \ln(0.5/0.5)\| + 0.5 \cdot 0.1\|\ln(0.1/0.5) - \ln(0.5/0.5)\|$	0.50

Table 5(b): Low Correlation Case, following Figure 7(b).

While log-likelihood LOG is a good baseline score for detecting outliers, it fails to detect some clear outliers, as shown in Figure 8. In that example the strength of the correlation among the attributes is the same in both normal and outlier examples and the only difference is in their feature distributions.

Normal

Feature1

0	0.5
1	0.5

Feature2

0	0	0.9
0	1	0.1
1	0	0.1
1	1	0.9

$$Log(D_{normal}, B, \theta) = -(0.5 \times \log 0.5 + 0.5 \log 0.5 + 0.5 \times 0.9 \log 0.9 + 0.5 \times 0.1 \log 0.1 + 0.5 \times 0.1 \log 0.1 + 0.5 \times 0.9 \log 0.9) = 0.44$$

$$FD_{Normal} = 0.5 \left| \log \frac{0.5}{0.5} \right| + 0.5 \left| \log \frac{0.5}{0.5} \right| + 0.5 \left| \log \frac{0.5}{0.5} \right| + 0.5 \left| \log \frac{0.5}{0.5} \right| = 0$$

Outlier

0	0.9
1	0.1

0	0	0.9
0	1	0.1
1	0	0.1
1	1	0.9

$$Log(D_{outlier}, B, \theta) = -(0.9 \log 0.5 + 0.1 \log 0.5 + 0.9 \times 0.9 \log 0.9 + 0.9 \times 0.1 \log 0.1 + 0.1 \times 0.1 \log 0.1 + 0.1 \times 0.9 \log 0.9) = 0.44$$

$$FD_{outlier} = 0.5 \left| \log \frac{0.9}{0.5} \right| + 0.5 \left| \log \frac{0.1}{0.5} \right| + 0.5 \left| \log \frac{0.5}{0.5} \right| + 0.5 \left| \log \frac{0.5}{0.5} \right| = 0.46$$

Fig. 8 An example of normal and outlier individuals and their conditional probability tables created using the Bayesian network shown in Figure 7(c). Log-likelihood assigns the same score to the normal and individuals in this example, while FD is able to differentiate between these two individuals.

Conversely, if there is a correlation among the attributes of individuals, the feature divergence score FD fails to take it into account and therefore fails to differentiate between normal and outlier individuals. Figure 9 shows an example where the normal individual has a correlation among its attributes while the outlier object does not have a correlation. The FD metric cannot separate those two individuals.

6 Theoretical Analysis and Comparison

The mathematical analysis of this section relates the KLD and ELD divergences defined in the above sections to other well-known scores. Readers who wish to proceed to the empirical evaluation can omit this section without loss of continuity.

A well-known Taylor series argument shows that KLD can be approximated by Pearson's χ^2 statistic (Nielsen and Nock, 2014). We use a similar Taylor

Normal

Feature1

0	0.5
1	0.5

Feature2

0	0	0.9
0	1	0.1
1	0	0.1
1	1	0.9

$$FD_{Normal} = 0.5\,|\log\frac{0.5}{0.5}| + 0.5\,|\log\frac{0.5}{0.5}| + 0.5\,|\log\frac{0.5}{0.5}| + 0.5\,|\log\frac{0.5}{0.5}| = 0$$

$$LR(D_{normal}, B, \theta) = 0.5\log\frac{0.5}{0.5} + 0.5\log\frac{0.5}{0.5} +$$
$$0.5\times 0.9\log\frac{0.9}{0.9} + 0.5\times 0.1\log\frac{0.1}{0.1} + 0.5\times 0.1\log\frac{0.1}{0.1} + 0.5\times 0.9\log\frac{0.9}{0.9} = 0$$

Outlier

0	0.5
1	0.5

0	0	0.5
0	1	0.5
1	0	0.5
1	1	0.5

$$FD_{outlier} = 0.5\log|\frac{0.5}{0.5}| + 0.5\log|\frac{0.5}{0.5}| + 0.5\log|\frac{0.5}{0.5}| + 0.5\log|\frac{0.5}{0.5}| = 0$$

$$LR(D_{outlier}, B, \theta) = 0.5\times\log\frac{0.5}{0.5} + 0.5\log\frac{0.5}{0.5} +$$
$$0.5\times 0.9\log\frac{0.5}{0.9} + 0.5\times 0.1\log\frac{0.5}{0.1} + 0.5\times 0.1\log\frac{0.5}{0.1} + 0.5\times 0.9\log\frac{0.5}{0.9} = -0.17$$

Fig. 9 An example of normal and outlier individuals and their conditional probability tables created using Bayesian network shown in Figure 7(a). FD assigns the same score to the normal and individuals in this example, while LR is able to differentiate between these two individuals.

series approximation to compare our ELD divergence with other divergences such as KLD. The analysis shows that the dominant component of ELD is the total variation distance (TVD), also known as the L1-norm, whose statistical properties are well understood (Beirlant et al., 2001, 1994).

Our analysis uses the f-divergence framework, which unifies all the standard divergence definitions for probability distributions. The f-divergence analysis shows the generality of the cancellation phenomenon: the first-order Taylor series term vanishes for all f-divergences whose generator f is differentiable at $\lambda = 1$, which includes all standard f-divergences (including KLD), except for those that use absolute values.

The χ^2 approximation for f-divergences. To focus the mathematics on the essential insights, we discuss the case of divergences defined over a single discrete random variable V with possible values $v_1, \ldots, v_m$. The results carry over to the joint BN distributions over discrete variables considered in this paper, by applying the approximation to each parent state. Given two probability distributions P_1 and P_2 over the values of V, an f-divergence is of the form

$$I_f(P_1||P_2) \equiv \sum_{i=1}^{m} P_1(v_i) f\left(\frac{P_2(v_i)}{P_1(v_i)}\right) \tag{9}$$

where the **generator** $f : R \to R$ is a convex function such that $f(1) = 0$. The generator transforms the ratio of the two compared probabilities for each possible value; the f-divergence is the weighted average of the transformed ratios. Common divergences can be represented by choosing a suitable generator. For example, KLD is generated by $f(u) = -\ln(u)$, the χ^2 statistic by $f(u) = (1-u)^2/u$, and the TVD by $f(u) = |u-1|$ (Nielsen and Nock, 2014).

An f-divergence I_f can be approximated by replacing f with its Taylor series expansion around the point $u = 1$. Assuming that the derivatives $f^{(i)}(1)$ of any order exist, the Taylor series for f takes the form

$$f(u) = \sum_{l=0}^{\infty} \frac{f^{(l)}(1)}{l!}[u-1]^l,$$

where the $l = 0$ term is 0 because $f(1) = 0$. Substituting the Taylor series expansion into the f-divergence expression (9) yields an f-divergence expansion, *assuming that all derivatives of f exist*:

$$I_f(P_1||P_2) = \sum_{l=0}^{\infty} \frac{f^l(1)}{l!} \sum_{i=1}^{m} \frac{[P_2(v_i) - P_1(v_i)]^l}{P_1(v_i)^{l-1}}. \tag{10}$$

A detailed derivation is in the appendix. Equation (10) shows that if the generator f is infinitely differentiable, the f-divergence is a series of l-metrics, scaled by the derivates and by the P_1 probabilities. For statistical applications, usually a second-order expansion up to $l = 2$ provides a sufficiently close approximation. If the generator f is twice differentiable, the first-order term with $i = 1$ vanishes, and the remaining second-order term with $i = 2$ is a scaled χ^2 statistic:

Proposition 1 (Nielsen and Nock 2014) *Assume that the generator f for I_f is twice differentiable.*

1. *The $i = 1$ Taylor series term for I_f around $u = 1$ is 0.*
2. *The $i = 2$ Taylor series term for I_f around $u = 1$ is a scaled χ^2 divergence.*

Therefore the second-order Taylor series approximation to I_f is scaled χ^2-divergence:

$$I_f(P_1||P_2) \approx \frac{f''(1)}{2}\chi^2(P_1||P_2) = \frac{f''(1)}{2} \sum_{i=1}^{m} \frac{(P_2(v_i) - P_1(v_i))^2}{P_1(v_i)}.$$

Proof Outline. The first-order term reduces to $f'(1)[\sum_{i=1}^m P_2(v_i) - \sum_{i=1}^m P_1(v_i)]$, which vanishes because both probability measures sum to 1. The second-order term reduces to $\frac{f''(1)}{2}\chi^2(P_1||P_2)$. The details are given in the Appendix.

The first clause provides a clear formulation of the cancelling phenomenon discussed in Section 5.2: the first-order terms cancel out exactly.

The Taylor series approximation for ELD. Our ELD metric is an example of transforming an f-divergence with absolute values by replacing the generator f with $|f|$; in our case, replacing $-\ln(u)$ by $|-ln(u)|$ (cf. Equation (6)). Because $|x|$ is not differentiable at 0, the absolute value generator $|f|$ is generally not differentiable at $u = 1$, so Proposition 1 does not apply. To utilize Taylor series analysis, we can rewrite the absolute value divergence as a positive sum where $f(u) > 0$ and a negative sum where $f(u) < 0$.

$$I_{|f|}(P_1||P_2) \equiv \sum_{i:f(\frac{P_2(v_i)}{P_1(v_i)})>0} P_1(v_1)f\left(\frac{P_2(v_i)}{P_1(v_i)}\right) - \sum_{i:f(\frac{P_2(v_i)}{P_1(v_i)})<0} P_1(v_1)f\left(\frac{P_2(v_i)}{P_1(v_i)}\right) \tag{11}$$

In the Taylor series approximation for *ELD* (using $f = -\ln(u)$ in Equation (11)), cancellation is avoided, so the first-order term does *not* vanish, and is in fact equivalent to the Total Variation Distance. The second-order term is equivalent to the **cancelled** χ^2 **metric**, which we define as

$$\overline{\chi}^2(P_1||P_2) \equiv 1/2 \sum_{i:P_2(v_i)<P_1(v_i)} [P_2(v_i{-}P_1(v_i]^2/P_1(v_i){-}1/2 \sum_{i:P_2(v_i)>P_1(v_i)} [P_2(v_i{-}P_1(v_i]^2/P_1(v_i)$$

Proposition 2 *Consider the Taylor series approximation for ELD around* $u = 1$ *with generator* $|f| = |-ln(u)|$.

1. *The* $i = 1$ *Taylor series term is is total variation distance:*
$$ELD(P_1||P_2) \approx TVD = \sum_i |P_2(v_i) - P_1(v_i)|$$
2. *The* $i = 2$ *Taylor series term for* I_f *around* $u = 1$ *is the cancelled* $\overline{\chi}^2(P_1||P_2)$ *divergence.*

In summary, a second-order Taylor series expansion shows that, up to constant factors, divergences can be approximated as first-order term corresponding to total variation distance (TVD), also known as the L1-norm, plus a second-order term corresponding to a χ^2 divergence. Cancellation occurs depending on the sign of the probability differences $P_1(v_i) - P_2(v_i)$: differences with opposing signs can cancel out. In KLD, complete cancellation occurs with the first-order term, so the L1-component vanishes. In our absolute value metric ELD, no first-order term is cancelled, but some second-order terms cancel, so the magnitude of the χ^2 is diminished.

7 Experimental Evaluation and Comparison

All the experiments were performed on a 64-bit Centos machine with 4GB RAM and an Intel Core i5-480 M processor. The likelihood-based outlierness scores were computed with SQL queries using JDBC, JRE 1.7.0. and MySQL Server version 5.5.34. We utilized the synthetic data sets and real-world data sets from the soccer, hockey, movie and biology domains.

Premier League Statistics		IMDB Statistics		NHL Statistics	
Number of Teams	20	Number of Movies	3060	Number of Teams	30
Number of Players	484	Number of Directors	220	Number of Players	921
Number of Matches	380	Number of Actors	98690	Number of Matches	720
avg player per match	26.01	avg actor per movie	36.42	avg player per match	29

Table 6 Summary Statistics for the IMDB and PL and NHL data sets

7.1 Synthetic Datasets

We generated three synthetic data sets with normal and outlier players using the distributions represented in the three Bayesian networks of Figure 7. The main goal of designing synthetic experiments is to test the methods on easy to detect outliers. We used the *mlbench* package in *R* to generate synthetic features in matches, following these distributions for 240 normal players and 40 outliers. (There were 280 players in our Premier League data set.) Each player participates in 38 matches, similar to the Premier League's real-world data. Each match assigns a value to each feature F_1 and F_2 for each player. This design yields the following three synthetic data sets.

High Correlation Normal individuals exhibit a strong association between their features, outliers no association. Both normals and outliers have a close to uniform distribution over single features. See Figure 7(a).

Low Correlation Normal individuals exhibit no association between their features, outliers have a strong association. Both normals and outliers have a close to uniform distribution over single features. See Figure 7(b).

Single Features Both normal and outlier individuals exhibit a strong association between their features. In normals, 90% of the time, feature 1 has value 1. For outliers, feature 1 has value 1 only 10% of the time. See Figure 7(c).

7.2 Real-world Datasets

Our data sets and code are available on-line (Riahi and Schulte, 2015b). The three real-world data sets we utilize are from the soccer, movie, and ice hockey domains. Table 6 shows summary statistics for the data sets. Table 7 lists their populations and features.

Premier League. The Opta data were released by Manchester City. It lists *box scores*: counts of all the ball actions within each game by each player, for the 2011-2012 season. For each player in a match, our data set contains eleven player features. For each team in a match, there are five features computed as player feature aggregates, as well as the team formation and the result (win, tie, loss). There are two relationships, $Appears\_Player(P, M)$, $Appears\_Team(T, M)$.

Individuals	Features
PL-Player per Match	$TimePlayed$, $Goals$, $SavesMade$, $ShotEff$, $PassEff$, $WinningGoal$, $FirstGoal$, $PositionID$, $TackleEff$, $DribbleEff$, $ShotsOnTarget$
PL-Team per Match	$Result$, $TeamFormation$, $\sum Goals$, $\mu ShotEff$, $\mu PassEff$, $\mu TackleEff$, $\mu DribbleEff$.
IMDB-Actor	$Quality$, $Gender$
IMDB-Director	$Quality$, $avgRevenue$
IMDB-Movie	$year$, $isEnglish$, $Genre$, $Country$, $RunningTime$, $Rating$ by User
IMDB-User	$Gender$, $Occupation$.
NHL-Player per Match	$Goals$, $Assists$, $Points$, $PowerPlayTime$, $PlusMinus$, $Penalties$, $Shots$, $Hits$, $BlockedShots$, $Giveaways$, $ShotHandedTime$, $TimeOnIce$
NHL-Team per Match	$Goals$, $GoalDifference$

Table 7 Attribute Features.

IMDb. The Internet Movie Database (IMDb) is an on-line database of information related to films, television programs, and video games. The IMDb website offers a data set containing information on cast, crew, titles, technical details and biographies into a set of compressed text files. We preprocessed the data like Peralta (2007) to obtain a database with seven tables: one for each population and one for the three relationships $Rated(User, Movie)$, $Directs(Director, Movie)$, and $ActsIn(Actor, Movie)$.

National Hockey League. The NHL provides information about the sequences of play-by-play events. We used the data crawled by Schulte and Routley (2014) comprising player game statistics for the seasons 2009-2013. The box scores summarize player statistics for each match, a total of 13 features. Following Schulte and Routley we use the total box scores over the player's most recent season as the player's feature vector.

Mutagenesis Data This data set contains mutagenic activity of 188 compounds. 125 of these compounds have positive levels of log mutagenicity that are labelled "active". The remaining compounds are labelled "inactive" and constitute the source of negative examples. In this data set we considered examples of active compounds as the normal population and the inactive ones as the outlier.

7.3 Evaluation Techniques for Outlier Detection

Measuring the effectiveness of outlier detection methods is easy. Most of the time ground truth information, that shows which data points are outliers, is unavailable. Several approaches have been taken in the literature to evaluate the performance of outlier detection methods:

1. Intuitive evaluation: case studies have been extensively used in the literature to evaluate outliers (Aggarwal, 2013). In Section 8.3 we use perform a case study on the top-n ranked outliers, by trying to explain and make sense of the detected outliers.
2. Synthetic data generation: another approach is generating synthetic data and injecting synthetic outliers into the data (Aggarwal, 2013). We have designed and developed three synthetic data sets as discussed in section 7.1.
3. Anomaly injection: anomalies are injected into real-world data sets. Outlier detection methods are expected to detect the injected data points as outliers (Akoglu et al., 2015). We employ this approach in our real-world data sets.

For anomaly injection in real-world data, we employ a one-class design: we learn a model for the class distribution, with data from that class only. Then we rank all individuals from the normal class together with all objects from a contrast class injected as outliers, and examine whether an outlierness score recognizes objects from the contrast class as outliers. *In-class outliers* are possible, meaning an object that is highly anomalous even within its class. For example, unusual strikers are still members of the striker class. In our unsupervised setting without class labels, we do not expect an outlierness score to distinguish such an in-class outlier from outliers outside the class. Our case studies describe a few in-class outliers.

Table 8 shows the normal and contrast classes for three different data sets. In the soccer data, we considered only individuals who played more than 5 matches out of a maximum 38. For the three synthetic data sets, the scores are used to rank all 280 synthetic players, 240 of which are normal and 40 are outliers.

7.4 Methods Compared

We compare three types of approaches, based on relational model likelihood, aggregation, and distance.

Normal class	#Normal	Outlier	#Outlier class
Striker	153	Goalie	22
Midfielder	155	Striker	74
Drama	197	Comedy	47
Defender	186	Forward	38
Positive Compound	125	Negative Compound	63

Table 8 Outlier/normal Objects in Real-World Datasets.

Likelihood-based Outlierness Scores. The first approach evaluates the likelihood-based outlierness scores described in Section 5. For relational Bayesian network structure learning we utilize the previous learn-and-join algorithm (LAJ), which is a state-of-the-art BN structure learning method for relational data (Schulte and Khosravi, 2012). The LAJ algorithm employs an iterative deepening strategy, which can be described as as search through a lattice of table joins. For each table join, different BNs are learned and the learned edges are propagated from smaller to larger table joins. For a full description, complexity analysis, and learning time measurements, please see (Schulte and Khosravi, 2012). We used the implementation of the LAJ algorithm due to its creators (Khosravi et al., 2019).

Aggregation-based Methods. The second approach first "flattens" the structured data into a matrix of feature vectors, then applies standard matrix-based outlier detection methods. We refer to such methods as **aggregation-based** (cf. Figure 2). For example, this was the approach taken by Breunig et al. (2000) for identifying anomalous players in sports data. Following their paper, for each continuous feature in the object data, we use the average over its values, and for each discrete feature, we use the occurrence count of each feature value in the object data. Aggregation tends to lose information about correlations. Our experiments address the question of *whether the loss of information through aggregation affects outlier detection.*

We applied three standard feature-based outlier detection methods: Density-based *LOF* (Breunig et al., 2000), distance-based *KNNOutlier* (Ramaswamy et al., 2000) and subspace analysis *OutRank* (Muller et al., 2012). These represent common representatives of fundamental approaches for vectorial data outlier detection. Like *ELD*, subspace analysis is sensitive to correlations among features. We used the available implementation of all three data matrix methods from the state of the art data mining software *ELKI* (Achtert et al., 2013). The clustering function for *OutRank* was *PRO-CLUS*, as recommended by (Muller et al., 2012).

Relational Distance-based Method. Our distance-based approach utilizes a first-order distance measure that was developed and for the instance-based learning system RIBL2 (Horváth et al., 2001). This measure has proven successful in several applications (Kirsten et al., 2001; Horváth et al., 1999). We compute RIBL2 distance between any two individuals in our population domain. Then, we rank the individuals based on their mean distance to the normal population.

8 Empirical Results

We present results regarding computational feasibility, predictive performance, and case studies.

8.1 Computational Cost of the *ELD* Score.

Table 9 shows that the computation of the *ELD* value for a given target object is feasible. On average, it takes a quarter of a minute for each soccer player, and one minute for each movie. This includes the time for parameter learning from the object database. Learning the class model BN takes longer, but needs to be done only once for the entire object class. This shows that *the BN model provides a compact low-dimensional representation of the joint distribution information in the data.*

Dataset	Class Model	Average per Object
PL: Strikers vs. Goalies	4.14	0.25
PL: Midfielder vs. Goalies	4.02	0.25
IMDb: Drama vs. Comedy	8.30	1.00
PL: Forward vs. Defender	5.30	0.35
Mutagenesis: Positive Compound vs. Negative Compound	1.40	0.1

Table 9 Time (min) for computing the *ELD* score.

8.2 Detection Accuracy

Several accuracy scores have been developed to measure the performance of outlier ranking methods (Aggarwal, 2013). Arguably the most widely used score is the *area under curve* (AUC) of the receiver operating characteristic ROC curve (Fawcett, 2006; Cansado and Soto, 2008; Muller et al., 2012). The relationship between false positive rate (1- Specificity) and true positive rate (Sensitivity) is captured by the ROC curve. The ideal performance is achieved at both the maximum sensitivity and the maximum specificity. The maximum value for AUC is 1.0 indicating a perfect ranking with 100% sensitivity and 100% specificity. In order to compute the AUC value, we used the R package $ROCR$ (Sing et al., 2012). Given a set of outlierness scores, one for each object, this package returns an AUC value.

AUC Accuracy in Synthetic Datasets. Table 10 shows the AUC values for probabilistic methods and the outlier detection methods. On the synthetic data, it is easy to distinguish the outliers and most methods do this well.

However, the EMM methods *LR* and *ELD* are the only scores that achieve perfect or near perfect detection across all three synthetic data sets.

RIBL is the only method that fails in detecting outliers. The reason is the following. *RIBL* computes the distance between two individuals as the minimum distance between two instances associated with the individuals (in their data profiles). For binary features, the distance between feature values is 0-1 depending on whether the features match. So in our experiment with two binary features, individuals whose profiles contain the same feature combinations receive distance 0. Since in our experiment, the distribution over feature combinations is different, but their support is the same. So almost all individuals are at distance 0 to each other and cannot be distinguished as outliers.

AUC Accuracy in Real-world Datasets. Table 11 shows the *AUC* values for aggregation-based methods compared to *ELD* and *RIBL* in the real-world data sets. *RIBL*'s recursion depth bound is set to 2. The top performance of *RIBL* and the model-based methods is substantially better than aggregation-based methods on all data sets, confirming that it is important to develop outlier detection methods based on relational statistics for the relational data. The EMM score *ELD* works better than *RIBL* or equally well in all the data sets except Midfielder vs. Strikers. In that data set *ELD* finds many in-class exceptional midfielders. Overall, our observations suggest that with a small number of discrete features in our data set, *RIBL* does compute a very informative distance and hence does not support outlier detection well. But for numeric features it seems to be a good alternative.

Comparing the two EMM methods, *ELD* achieves a substantially higher AUC than *LR* on all data sets except for the NHL. On the NHL data set, the *LR* metric outperforms both *RIBL* and *ELD*, which is an example of how taking absolute values does not alway result in better detection performance.

On the synthetic data, we compare the two EMM methods further with precision@r%, another metric used in a previous study on relational outlier detection (Gao et al., 2010), which is computed as follows: 1) Sort the outlierness scores obtained by an outlier ranking method in descending order. 2) Return the fraction of the top r percent that are in fact outliers. Precision@r% focuses on the ability to detect clear outliers. As Table 12 shows, the *ELD* score is the best for this metric.

Conclusion. The outlier detection methods with the best over all performance are the two based on the exceptional model mining framework, using the traditional log-likelihood ratio (Kulback-Leibler divergence), and our novel expected log-likelihood distance metric *ELD*. While the *ELD* score does not uniformly outperform *LR*, the improvement is frequent and substantial enough to recommend it as a default method for practical applications.

Dataset	<u>*ELD*</u>	*LOG*	<u>*LR*</u>	*FD*	*RIBL*	*LOF*	*OutRank*	*KNN Outlier*
High Correlation	**1.00**	0.99	0.97	0.89	0.50	0.68	0.99	0.97
Low Correlation	**1.00**	0.97	0.99	0.42	0.50	0.58	0.83	0.97
Single Feature	**1.00**	0.79	**1.00**	**1.00**	0.50	0.63	0.88	0.86

Table 10 AUC of the *ELD* Outlier detection methods in Synthetic data sets. The EMM methods are underlined.

Dataset	<u>*ELD*</u>	*LOG*	<u>*LR*</u>	*FD*	*RIBL*	*LOF*	*OutRank*	*KNN Outlier*
PL: Strikers vs. Goalies	**0.89**	0.61	0.65	0.61	0.71	0.61	0.60	0.61
PL: Midfielders vs. Strikers	0.66	0.45	0.55	0.59	**0.78**	0.76	0.71	0.58
IMDb: Drama vs. Comedy	**0.70**	0.66	0.66	0.64	**0.70**	0.51	0.68	0.68
NHL: Defender vs. Forward	0.78	0.58	**0.87**	0.79	0.71	0.73	0.73	0.66
Mutagenesis: Positive vs Negative	**0.86**	0.64	0.81	0.70	0.70	0.51	0.57	0.53

Table 11 AUC outlier detection methods in real-world data sets.

Table 12 Precision@r% of Outlier detection methods in Synthetic data sets.

Dataset	Percentage	<u>*ELD*</u>	*LOG*	<u>*LR*</u>	*FD*	*LOF*	*OutRank*	*KNN Outlier*
High-Correlation	1%	**1.00**	0.91	0.73	0.47	0.11	0.53	0.48
	5%	**1.00**	0.95	0.85	0.65	0.22	0.50	0.65
Low-Correlation	1%	**1.00**	0.93	0.87	0.14	0.10	0.00	0.06
	5%	**1.00**	0.95	0.90	0.25	0.25	0.10	0.57
Single-Feature	1%	**1.00**	0.81	**1.00**	**1.00**	0.46	**1.00**	0.51
	5%	**1.00**	0.89	**1.00**	**1.00**	0.55	**1.00**	0.54

8.3 Case Studies

For a case study, we examine the three top outliers as ranked by *ELD*, shown in Table 13. The aim of the case study is to provide a qualitative sense of the outliers indicated by the scores. Also, we illustrate how the BN representation leads to an interpretable ranking. Specifically, we employ a *feature-wise decomposition* of the score combined with a *drill down* analysis:

1. Find the node V_i that has the highest ELD_i divergence score for the outlier object.
2. Find the parent-child combination that contributes the most to the ELD_i score for that node.
3. Decompose the ELD_i score for the parent-child combination into feature and mutual information component.

We present strong associations—indicated by the *ELD*'s mutual information component—in the intuitive format of association rules. Association analysis can also be applied to *LR* using the form (5); we focus on *ELD* for brevity.

Strikers vs. Goalies. Edin Dzeko is highly anomalous striker who obtains the top *ELD* outlierness score among both strikers and goalies. His *ELD* score is highest for the Dribble Efficiency feature. The highest *ELD* score for that feature occurs when Dribble Efficiency is low, and its parents have the following values: Shot Efficiency = high, Tackle Efficiency = medium. Looking at the single feature divergence, we see that Edin Dzeko is indeed an outlier in the Dribble Efficiency subspace: His dribble efficiency is low in 16% of his matches, whereas a randomly selected striker has low dribble efficiency in 50% of their matches. Thus, Edin Dzeko is an unusually good dribbler. Fans have recognized his skill by posting many examples of Dzeko dribbling on Youtube. Looking at the mutual information component of *ELD*, for Edin Dzeko the confidence of the rule

$$\mathit{ShotEff} = \mathit{high}, \mathit{TackleEff} = \mathit{medium} \rightarrow \mathit{DribbleEff} = \mathit{low}$$

is 50%, whereas in the general striker class it is 38%. Therefore his Shot Efficiency and Tackle Efficiency have an unusually high association with his Dribble Efficiency.

Soccer Midfielders vs. Strikers. For the single feature score, Robin van Persie is recognized as a clear striker because of the *ShotsOnTarget* feature. It makes sense that strikers shoot on target more often than midfielders. Robin van Persie achieves a high number of shots on targets in 34% of his matches, compared to 3% for a random midfielder. The mutual information component shows that he also exhibits unusual correlations. For example, the confidence of the rule

$$\mathit{ShotEff} = \mathit{high}, \mathit{TimePlayed} = \mathit{high} \rightarrow \mathit{ShotsOnTarget} = \mathit{high}$$

is 70% for van Persie, whereas for strikers overall it is 52%. The most anomalous midfielder is Scott Sinclair. His most unusual feature is *DribbleEfficiency*: For feature divergence, he achieves a high dribble efficiency 50% of the time, compared to a random midfielder with 30%.

Drama vs. Comedy. The top *ELD* rank is assigned to the in-class outlier *BraveHeart*. Its most unusual feature is *ActorQuality*: In a random drama movie, 42% of actors have the highest quality level 4, whereas for *BraveHeart* 93% of actors achieve the highest quality level. The *ELD* score identifies the comedies *BluesBrothers* and *AustinPowers* as the top out-of-class outliers. In a random drama movie, 49% of actors have casting position 3, whereas for *AustinPowers* 78% of actors have this casting position, and for *BluesBrothers* 88% of actors do.

Hockey Defenders vs. Forwards. The first two players in the ranking are Forwards; the high outlierness ranking is mainly due to their unusually high value for points. Eric Staal has $\mathit{points} = 2$ in 30% of his matches, while a random player has that value for points only on 6% of his matches. Dustin Byfuglien is an in-class outlier. His most unusual feature is *PowerPlayTime*. While an average player shows $\mathit{PowerPlayTime} = 2$ for only 33% of his matches, Byfuglien's percentage is 79%.

Strikers (Normal) vs. Goalies (Outlier)					
PlayerName	Position	*ELD* Rank	*ELD* Max Node	*ELD* Node Score	*FD* Max feature Value
Edin Dzeko	Striker	1	DribbleEfficiency	83.84	DE=low
Paul Robinson	Goalie	2	SavesMade	49.4	SM=Medium4
Michel Vorm	Goalie	3	SavesMade	85.9	SM=Medium
Midfielders (Normal) vs. Strikers (Outlier)					
PlayerName	Position	*ELD* Rank	*ELD* Max Node	*ELD* Node Score	*FD* Max feature Value
Robin Van Persie	Striker	1	ShotsOnTarget	153.18	ST=high
Wayne Rooney	Striker	2	ShotsOnTarget	113.14	ST=high
Scott Sinclair	Midfielder	6	DribbleEfficiency	71.9	DE=high
Drama (Normal) vs. Comedy (Outlier)					
MovieTitle	Genre	*ELD* Rank	*ELD* Max Node	*ELD* Node Score	*FD* Max feature Value
Brave Heart	Drama	1	ActorQuality	89995.4	a_quality=4
Austin Powers	Comedy	2	Cast_Position	61021.28	Cast_Num=3
Blue Brothers	Comedy	3	Cast_Position	24432.21	Cast_num=3
Defender (Normal) vs. Forward (Outlier)					
PlayerName	Position	*ELD* Rank	*ELD* Max Node	*ELD* Node Score	*FD* Max feature Value
Eric Staal	Forward	1	Points	49.57	Points=2
Phil Kessel	Forward	2	Points	43.34	Points=2
Dustin Byfuglien	Defender	3	Power_PlayTime	25.65	PP_time=2

Table 13 Case study for the top outliers returned by the *ELDscore*

9 Correlation with Success

The aim of this section is to compare the *ELD* metric with other meaningful metrics for comparing individuals. Our reference metrics are *success rankings* of individuals selected for a specific domain, shown in Table 14. We use the same data as in our other experiments, described in Section 7. (We left out Mutagenesis as there is no meaningful success metrics for that data set.)

Success rankings are one of the most interesting features to users. Strong correlations between the *ELD* metric and meaningful success metrics provide evidence that the *ELD* metric is meaningful as well. We measure correlation strength by the standard Pearson correlation coefficient ρ. The coefficient ranges from -1 to 1, where 0 means no correlation and 1 or -1 indicates maximum strength (Fisher, 1921).

The observed correlations are remarkable in at least two respects.

1. The strength of the correlations are high. For example, between *ELD* and salary, coefficients range from 0.45 to 0.82 (see Table 16).
2. We observe high correlations across different domains, different types of individuals, and different success metrics.

For a population with a diverse set of skills and resources, being different from the generic class can be interpreted as both exceptionally better or worse than the normal population. In the domains we study in this data, we found that higher *ELD* scores point to exceptionally good individuals but not to exceptionally bad individuals. Our interpretation of the correlation between *ELD* and success, rather than failure, is that our domains featured skilled individuals, so a normal individual is quite successful already. For example, in the Premier League we expect most players to be in the range of good players. Therefore, deviating from the rest of the population is a signal for detecting

Dataset	Success Metric	Min	Max	Standard Dev.	Mean
IMDb	Sum of Rating	1.0	14795	1600.22	1057.58
PL-Player	TimePlayed	5.0	3420	1015.69	1484.00
PL-Player	Normalized Salary	0.007	0.28	0.62	0.10
PL-Player	Sum of Shot Efficiency	0	82	9.87	6.53
PL-Team	Standing	1.0	20	5.91	10.50
NHL-Player	Power Play Time	0	669	106.78	84.38
NHL-Player	Time on Ice	4	2099	278.03	1187.31
NHL-Player	Assists	0	4	0.49	0.20

Table 14 Success metrics and their distributions.

exceptionally good players. Our *ELD*-success scatterplots below (Figure 11) provide empirical evidence for this interpretation: we typically see a large cluster of individuals around the origin, meaning that their success level is normal and their *ELD* score is low.

Discussion of Success Metrics. Sum of ratings for movies and league standing for teams are immediate measures of success. Measures like time played and shot efficiency are proxies for success that are widely used in sports analytics. Recent work in sports analytics has introduced sophisticated new performance metrics with several advantages. For example, time played is sensitive to factors beyond a player's control, such as injuries and coaching tactics. By contrast, the goal impact metric, which is based on reinforcement learning, focuses on the value of the individual actions a player performs (Routley and Schulte, 2015; Liu and Schulte, 2018).

Our outlierness metrics are computed from instantiation counts (e.g. total passes in a match). However, their definition does not entail that they are proportional to such counts, for two reasons: i) The metrics are normalized to frequencies (e.g. passes per match, not total passes over all matches). ii) The metrics measure not directly the frequencies, but the *difference in frequencies* between a random and a specific individual. The correlation between our outlierness metrics and count-based success measures can therefore not be explained as a matter of definition only.

9.1 Methodology

We report the correlations between the *ELD* metric and metrics of success for a specific domain. We also focus on some unusually successful individuals as case studies. In considering the correlation between *ELD* and success, it is useful to investigate subgroups of individuals to ensure an apples-to-apples comparison (Sun et al., 2009). For instance, the attributes that lead to success are different for strikers and goalies. Accordingly, we report correlations for subgroups as well as entire classes of individuals. We found that compared to other model-based outlierness metrics, *ELD* shows the strongest correlation

Team	Standing
Top Teams	-0.71
Bottom Teams	-0.33
All Teams	-0.13

Table 15 Correlation between *ELD* metric and standing of Teams. The best standing is place 1.

Class	Time Played	Salary	Saves Made	Shots On target	Pass Efficiency
Strikers	0.86	0.82	NA	0.79	NA
Midfielders	0.80	0.45	NA	NA	0.77
Goalies	0.77	NA	0.74	NA	NA
All players	0.18	0.56	NA	NA	NA

Table 16 Correlation between *ELD* metric and success metrics of Soccer Players.

Genre	Sum of Rating	Average of Rating	Number of Rating
Action	0.67	0.30	0.72
Drama	0.76	0.29	0.81
Comedy	0.85	0.41	0.84
All Movies	0.56	0.17	0.60

Table 17 Correlation between *ELD* metric and success metric of Movies.

with success for all metrics and subgroups, except for the log-likelihood metric; detailed comparisons follow in Section 9.3.

9.2 Correlations between the *ELD* outlier metric and success

The next three tables summarize the observed correlations between success and *ELD* metrics: Teams in Table 15, Players in Table 16, Movies in Table 17.

9.2.1 Soccer Teams

Team Standing. The most successful team has Standing=1 and the least successful team has Standing=20 in the 2011-2012 Season. Figure 10 shows the correlation of *ELD* with team success metrics in a scatter plot. For the top teams, a very strong negative correlation emerges between *ELD* and standing: teams with higher *ELD* achieve a better (lower) standing. The top two teams Manchester City and Manchester United stand out very strongly in terms of the *ELD* metric (bottom right corner).

9.2.2 Soccer Players

Players Time Played is the total time that a player played over all matches in the season. This metric was shown to correlate strongly with other success

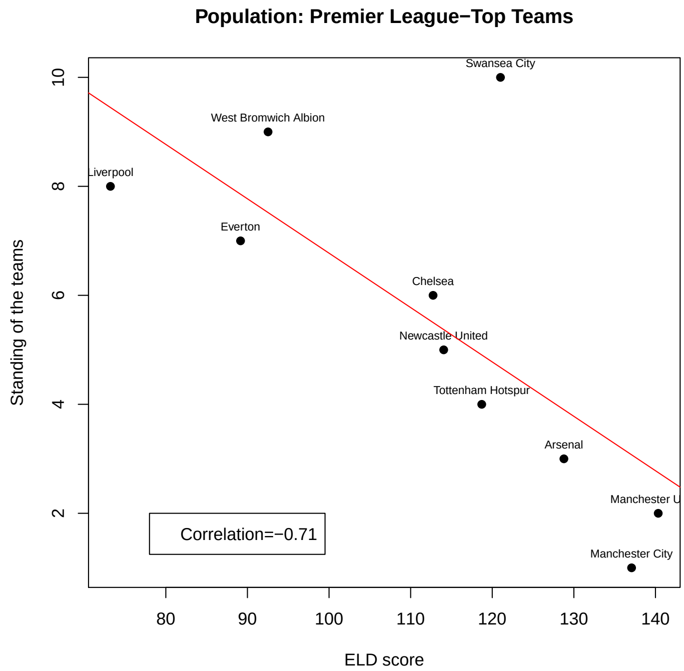


Fig. 10 Teams: Team Standing vs. *ELD* for the top teams in Premier League.

metrics, such as salary, in soccer data (Albert et al., 2017). For each subgroup, there is a strong positive correlation with *ELD*, meaning that atypical players with higher *ELD* tend to play more minutes.

Salary is probably the most obvious, and at the same time often the most misleading way to measure success of the players. Previous studies suggest that salary of the players does not always follow their performance in many sports such as Baseball and Soccer (Hall et al., 2002; Garcia-del Barrio and Pujol, 2004). They show that pay cannot be explained only by past performance but also depends on other factors that are hard to quantify.

We manually collected salaries of 120 players from internet sources. Table 16 and Figures 11 and 12 show the correlation between *ELD* and this success metric. The correlation is high, especially for Strikers. We found salary data for only 5 goalies. We discuss the relatively weaker salary correlation for midfielders in more detail below.

Shots on Target applies to strikers only. This is defined as any shot attempt that would or does enter the goal if left unblocked. We record the total number of these shots over all matches of the strikers only. This metric was shown to correlate strongly with *ELD* (see Table 16, Figure 11(b)). Figure 11 plots *ELD* against striker success metrics summed over the entire 2011-2012 season. We observe a large cluster around the origin, which represents a large base of normal strikers with average salaries and low *ELD* scores. The striker with the greatest *ELD* score is Robin van Persie. He stands out in terms of Shots on Target, Time Played, and Salary.

Saves Made applies to Goalies only. It is defined as the total number of saves that goalies had made over all the matches. This metric shows a strong correlation with *ELD* as well (see Table 16, Figure 12(b)). Figure 12 shows a scatter plot with the correlation of *ELD* with Goalie success metrics, summed over the entire 2011-2012 season. Goalies do not vary much in terms of the time they play. Wayne Hennessey has the highest number of Saves Made and also an unusually high *ELD* score, although not the highest.

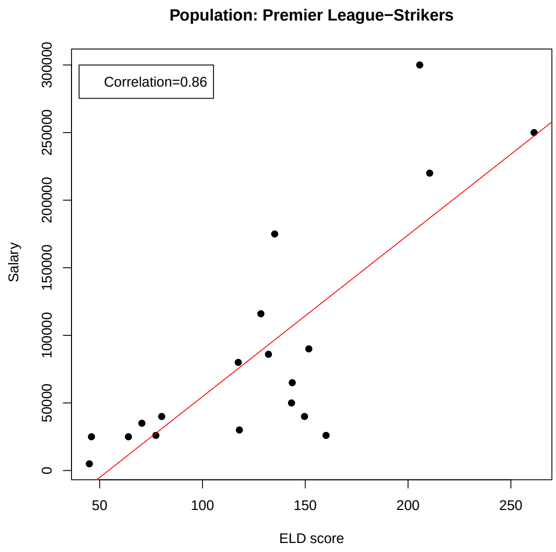
Midfielder Salary. We examine midfielder salaries for the 2011-2012 season. We omit a scatterplot for midfielder salary vs. *ELD* because it is less informative due to the weaker correlation (0.45). To investigate the reason for the weaker correlation, we picked two midfielders: 1) Stephane Sessegnon who has been ranked second in the *ELD* ranking but does not draw a large salary. 2) Steven Gerrard who is a very well known player and ranked second in the Salary ranking but according to the *ELD* score, he has been ranked 21. Based on domain knowledge, we picked some of the features that are relevant to midfield performance from the raw data and compared the feature statistics for these two players. Table 18 shows the details of their appearances in different matches. Sessegnon scored higher than Gerrard in three out of the four categories (Passes and Time Played). However, his salary was much lower than Gerrard's. Indeed his next contract with West Bromwich Albion netted him a club record fee. This is an example of how our *ELD* metric can identify players who are underpaid relative to their potential.

Name	Team	age	Salary Ranking	*ELD* Ranking	Time Played	Unsuccessful Passes	Successful Long Passes	Successful corners
Steven Gerrard	Liverpool	31	2	21	1212 min	244	52	25
Stephane Sessegnon	Sunderland	26	22	2	3133 min	231	82	15

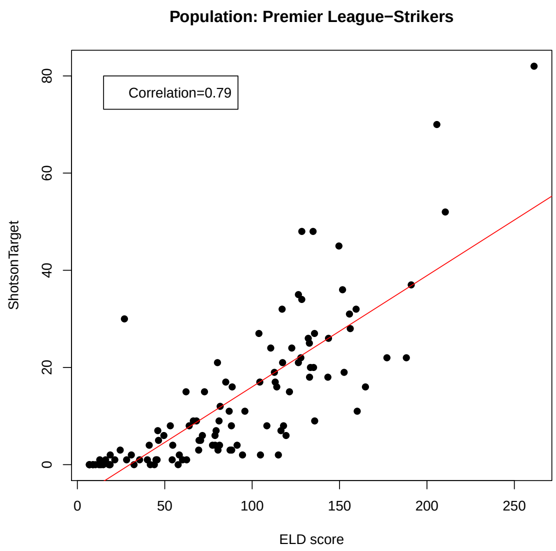
Table 18 Comparison of two midfielders.

Class	Time On Ice	Power Play Time	Assists	Goals
Forwarder	0.79	0.78	0.81	0.78
Defender	0.66	0.57	0.60	0.40

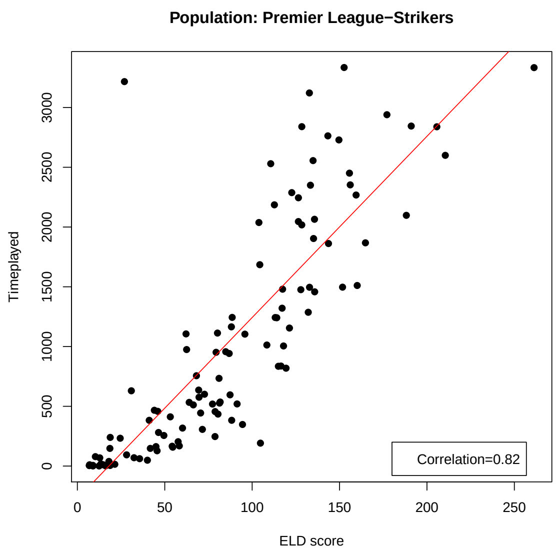
Table 19 Correlation between *ELD* metric and success metrics of NHL Players.



(a) Strikers: Salary vs *ELD*.

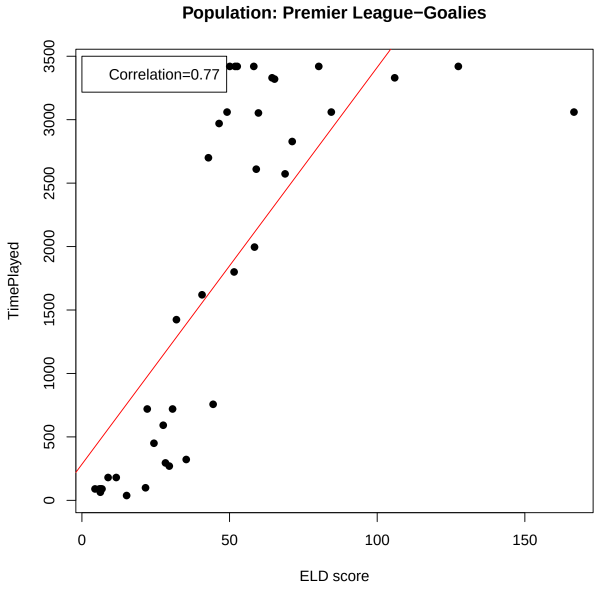


(b) Strikers: Shots On Target vs *ELD*

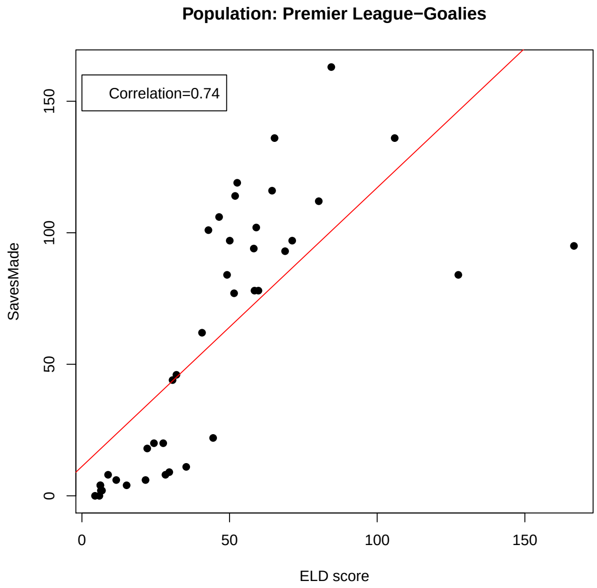


(c) Strikers: Time played vs *ELD*.

Fig. 11 Correlations between *ELD* and success metrics for strikers.



(a) Goalies: sum of time played vs *ELD*.



(b) Goalies: sum of saves made vs *ELD*

Fig. 12 Correlations between *ELD* and success metrics for Goalies

9.2.3 Hockey Players

For the hockey players the features that we have selected as success metrics are as follows: *TimeonIce* is the total amount of time a player has played over the course of a season. *PowerPlayTime* is the total amount of power player time for a player in a season. *Assists* is the total number of any actions by a player that led to a goal. Finally we use the total number of *Goals* that a player has scored. The correlation between these features and *ELD* is shown in Table 19. ELD has high correlations in features of both categories, however, it is stronger in the Forward subgroup.

9.2.4 Movies

Movie Sum of Ratings is the number of user ratings of a movie. Table 17 shows a high correlation with the *ELD* metric. The highest correlation obtains for the genre Comedy (0.84). The correlation between a movie and the sum of its ratings is equally strong, but the correlation with its average rating is much weaker. Thus the *ELD* score is associated with how many users have rated the movie rather than with how they have rated it. The number of ratings is a meaningful success metric, as it measures the number of people who have gone to see the movie.

9.3 Correlations between other outlierness metrics and success

In Section 3.4 and 5 we introduced other metrics that could be used in order to detect outliers. In this section we discuss the correlation between those metrics and success. The correlation between *ELD* and success is always stronger than *FD* and *LR* in all the data sets and subgroups, therefore we omit those results. The Log-likelihood is the only metric that results in a stronger correlation in some data sets and subgroups. Table 20 shows the subgroups for which log-likelihood shows a stronger correlation than the *ELD* metric. A high correlation with success can be interpreted as detecting in-class outliers. This results shows that log-likelihood could be an alternative score for finding in-class outliers.

10 Limitation of Model-based outlier detection

Quantitative evaluation, case studies, and correlation with meaningful success metrics confirm the effectiveness of exceptional model mining as an outlier detection method for relational data. The main limitations of our approach are the following.

1. Our proposed method ranks potential outliers, but does not set a threshold for a binary identification of outlier vs. non-outlier.

Subgroup	Success metric	*ELD* correlation	Log-likelihood correlation
Comedy	Sum of Rating	0.85	0.87
Drama	Sum of Rating	0.78	0.82
PL-Midfielder	Pass Efficiency	0.77	0.89
PL-Midfielder	Time Played	0.80	0.86
PL-Goalies	Time Played	0.77	0.87
PL-Goalies	Saves Made	0.74	0.85
NHL-Forward	Time on Ice	0.79	0.95
NHL-Forward	Goals	0.78	0.83

Table 20 Subgroups and success metrics for which the log-likelihood metric's correlates with success stronger than *ELD*.

2. Our current Bayesian network learning method can only be applied to discrete data. Prior to learning the model, we converted continuous data into discrete, which causes some information loss.
3. Our generative model-based methods learn a generic Bayesian network structure for the entire population, so the detected outliers are global outliers. However, there are more complex outliers that locally deviate from their subgroups and can be detected only by subgroup comparison. One direction for future work is to first detect subgroups in the population and then outliers within subgroups.[2]

Another limitation of our work is that, we used only part of the full information available in our rich data sets. Model-based outlier detection can be extended to take advantage of the full information, in the following manner.

1. In the Premier League data set, players are naturally related to one another and modeling the interaction between players can be another way to detect anomalous players. Graph-based features, such as detecting near-clique nodes and star nodes, proved to be informative for anomaly detection as shown in the ODDBALL system (Akoglu et al., 2010).
2. In this paper we did not use the temporal information available in the data. In the learning process we do not give a higher weight to more recent observations. This point is especially important when applying the methods to dynamic data or the data that are collected over long periods of time.
3. We did not incorporate ways to estimate missing values in model learning and/or outlierness scores.

11 Conclusion and Future Work

We presented a new approach for applying Bayesian networks to object-relational outlier detection, a challenging and practically important topic for machine learning. The key idea is to apply the exceptional model mining framework as follows. First, use statistical-relational learning to construct from relational

[2] We are indebted to Daniel Lowd for this point.

data a graphical model. Then learn one set of parameter values that represent class-level associations, another set to represent object-level associations, and compare how well each parametrization fits the relational data that characterize the target object. The classic metric for comparing two parametrized models is their likelihood ratio. As an novel alternative, we define a new relational outlierness score via two transformations: (1) a mutual information decomposition, and (2) replacing log-likelihood differences by log-likelihood distances. This score combines a single feature component, where features are treated as independent, with a correlation component that measures the deviation in the features' mutual information.

In experiments on three synthetic and four real-world outlier sets, the EMM methods based on log-likelihood difference and the log-likelihood distance achieved the best detection accuracy. On all but one real-world data set, log-likelihood distance outperformed the log-likelihood difference. As an alternative to model-based EMM, converting the structured data to a flat data matrix via aggregation had a negative impact on outlier detection. Case studies showed that the EMM scores lead to easily interpreted rankings. We found that the expected log-likelihood distance score correlated with success metrics to a surprising degree, across different domains and classes of individuals. The correlation with metrics of independent interest corroborates that this outlierness score produces meaningful and interesting results.

There are several avenues for future work. (i) A limitation of our current approach is that it ranks potential outliers, but does not set a threshold for a binary identification of outlier vs. non-outlier. (ii) Our divergence uses expected L1-distance for interpretability, but other metrics like L2 could be investigated as well. (iii) Extending the expected L1-distance for continuous features is a useful addition. (iv) Comparing our metric with the interestingness measures that have been developed for relational exception mining would increase our understanding of different how individuals can be unusual in structured data. (v) In the movie and soccer domains, our metric identified exceptionally successful individual objects, but not exceptionally unsuccessful ones. Our hypothesis was that in these domains, individuals have gone through a rigorous selection process, so the normal baseline performance is high. While we provided evidence for this hypothesis, it can be further investigated, by applying our outlier detection to data sets that feature a range of skills, including both amateur and professional performance. (vi) The theoretical properties of our *ELD* metric are important to understand. For example, when is the Taylor series approximation of *ELD* by Total Variation Distance sufficient close that theoretical guarantees for TVD hold also for *ELD*?

In sum, outlier metrics based on model likelihoods are a new type of structured outlierness score for object-relational data. Our evaluation indicates that model-based scores provide informative, interpretable, and accurate rankings of objects as potential outliers.

Acknowledgements

This work was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada. We are indebted for helpful discussions to the participants at the 2018 Workshop on Statistical Relational AI. The action editor and reviewers for the Data Mining and Knowledge Discovery Journal provided extensive helpful comments. We thank Peter Flach for clarifying the connection with exceptional model mining.

Appendix: Proofs

We prove the results about f-divergences from Section 6. We first review the known result that twice differentiable f-divergences are approximately scaled χ^2-divergences. Second we prove our new result that our new *ELD* metric is approximately total variation distance (plus a quadratic term with small impact).

Assume that the generator f is twice differentiable for f-divergence I_f. Then $I_f(P_1||P_2) \approx \frac{f''(1)}{2}\chi^2(P_1||P_2) = \frac{f''(1)}{2}\sum_{i=1}^{m}\frac{(P_2(v_i)-P_1(v_i))^2}{P_1(v_i)}$.

Proof Truncating the Taylor series for f at $i = 2$, the f-divergence expression (9) becomes

$$\begin{aligned}
&\sum_{i=1}^{m} P_1(v_i)[f'(1)\left(\frac{P_2(v_i)}{P_1(v_i)}-1\right)+\frac{f''(1)}{2}\left(\frac{P_2(v_i)}{P_1(v_i)}-1\right)^2] \\
&=\sum_{i=1}^{m} P_1(v_i)[f'(1)\left(\frac{P_2(v_i)-P_1(v_i)}{P_1(v_i)}\right)+\frac{f''(1)}{2}\left(\frac{P_2(v_i)-P_1(v_i)}{P_1(v_i)}\right)^2] \\
&=f'(1)[\sum_{i=1}^{m} P_2(v_i)-\sum_{i=1}^{m} P_1(v_i)]+\sum_{i=1}^{m}\frac{f''(1)}{2}\frac{(P_2(v_i)-P_1(v_i))^2}{P_1(v_i)} \\
&=0+\frac{f''(1)}{2}\chi^2(P_1||P_2)
\end{aligned}$$

The first-order Taylor series approximation for ELD is total variation distance:

$$ELD(P_1||P_2) \approx TVD.$$

Proof Truncating the Taylor series for $f = -\ln(u)$ at $i = 1$, the $|f|$-divergence expression (11) becomes

$$
\begin{aligned}
&\sum_{i:f(\frac{P_2(v_i)}{P_1(v_i)})>0} P_1(v_i)f'(1)\left(\frac{P_2(v_i)}{P_1(v_i)}-1\right) - \sum_{i:f(\frac{P_2(v_i)}{P_1(v_i)})<0} P_1(v_i)f'(1)\left(\frac{P_2(v_i)}{P_1(v_i)}-1\right)\\
&= \sum_{i:f(\frac{P_2(v_i)}{P_1(v_i)})>0} P_1(v_i)f'(1)\left(\frac{P_2(v_i)-P_1(v_i)}{P_1(v_i)}\right) - \sum_{i:f(\frac{P_2(v_i)}{P_1(v_i)})<0} P_1(v_i)f'(1)\left(\frac{P_2(v_i)-P_1(v_i)}{P_1(v_i)}\right)\\
&= -\sum_{i:P_2(v_i)<P_1(v_i)} P_2(v_i)-P_1(v_i) + \sum_{i:P_2(v_i)>P_1(v_i)} P_2(v_i)-P_1(v_i) \qquad (12)\\
&= \sum_i |P_2(v_i)-P_1(v_i)|
\end{aligned}
$$

Equation (12) holds for $f = -\ln$ because then $f(\frac{P_2(v_i)}{P_1(v_i)}) > 0$ if and only if $P_2(v_i) < P_1(v_i)$ and also $f'(1) = -1$.

References

Elke Achtert, Hans-Peter Kriegel, Erich Schubert, and Arthur Zimek. Interactive data mining with 3d-parallel coordinate trees. In *Proceedings ACM Special Interest Group on Management of Data*, pages 1009–1012, New York, NY, USA, 2013. doi: 10.1145/2463676.2463696. URL http://doi.acm.org/10.1145/2463676.2463696.

Charu C Aggarwal. *Outlier Analysis*. Springer New York, 2013. ISBN 9781461463955. URL http://books.google.ca/books?id=900CkgEACAAJ.

Leman Akoglu, Mary ~~Mcglohon~~, and Christos Faloutsos. Oddball: Spotting anomalies in weighted graphs. In *Proceedings Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 410–421, 2010. doi: 10.1007/978-3-642-13672-6_40. URL https://doi.org/10.1007/978-3-642-13672-6{\%}5C{_}40.

Leman Akoglu, Hanghang Tong, and Danai Koutra. Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery*, 29(3):626–688, 2015.

Jim Albert, Mark E Glickman, Tim B Swartz, and Ruud H Koning. *Handbook of Statistical Methods and Analyses in Sports*. CRC Press, 2017.

Grant Anderson and Bernhard Pfahringer. Exploiting propositionalization based on random relational rules for semi-supervised learning. In *Proceedings Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 494–502, 2008. doi: 10.1007/978-3-540-68125-0_43. URL http://dx.doi.org/10.1007/978-3-540-68125-0{_}43.

Fabrizio Angiulli, Gianluigi Greco, and Luigi Palopoli. Outlier detection by logic programming. *ACM Transactions on Computational Logic*, 9(1(7)):7, 2004.

Jan Beirlant, László Györfi, and Gábor Lugosi. On the asymptotic normality of the L1-and L2-errors in histogram density estimation. *Canadian Journal of Statistics*, 22(3):309–318, 1994.

Jan Beirlant, Luc Devroye, László Györfi, and Igor Vajda. Large deviations of divergence measures on partitions. *Journal of Statistical Planning and Inference*, 93(1-2):1–16, 2001.

Markus Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: Identifying density-based local outliers. In *Proceedings ACM Special Interest Group on Management of Data*, pages 93–104, 2000. doi: 10.1145/342009.335388. URL `https://doi.org/10.1145/342009.335388`.

Antonio Cansado and Alvaro Soto. Unsupervised anomaly detection in large databases using Bayesian networks. *Applied Artificial Intelligence*, 22:309–330, 2008. ISSN 0883-9514. doi: 10.1080/08839510801972801.

Luis de Campos. A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. *JMLR*, 7:2149–2187, 2006. URL `http://jmlr.org/papers/v7/decampos06a.html`.

Pedro Domingos and Daniel Lowd. *Markov Logic: An Interface Layer for Artificial Intelligence*. Morgan and Claypool Publishers, 2009.

Pedro Domingos and Matthew Richardson. Markov logic: A unifying framework for statistical relational learning. In *Introduction to Statistical Relational Learning*. MIT Press, 2007.

Wouter Duivesteijn, Ad J Feelders, and Arno Knobbe. Exceptional model mining. *Data Mining and Knowledge Discovery*, 30(1):47–98, 2016.

Tom Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006. doi: 10.1016/j.patrec.2005.10.010. URL `https://doi.org/10.1016/j.patrec.2005.10.010`.

Ronald A Fisher. On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32, 1921.

Jing Gao, Feng Liang, Wei Fan, Chi Wang, Yizhou Sun, and Jiawei Han. On community outliers and their efficient detection in information networks. In *Proceedings ACM SIGKDD*, KDD '10, pages 813–822, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0055-1. doi: 10.1145/1835804.1835907. URL `http://doi.acm.org/10.1145/1835804.1835907`.

Pedro Garcia-del Barrio and Francesc Pujol. Pay and performance in the ~~spanish~~ soccer league: Who gets the expected monopsony rents? Faculty Working Papers 05/04, School of Economics and Business Administration, University of Navarra, March 2004. URL `https://ideas.repec.org/p/una/unccee/wp0504.html`.

Lise Getoor. *Learning Statistical Models From Relational Data*. PhD thesis, Department of Computer Science, Stanford University, 2001.

Lise Getoor and Ben Taskar. *Introduction to Statistical Relational Learning*. MIT Press, 2007.

Stephen Hall, Stefan Szymanski, and Andrew S Zimbalist. Testing Causality Between Team Performance and Payroll: The Cases of Major League Baseball and English Soccer. *Journal of Sports Economics*, 3(2):149–168, 2002. URL `http://jse.sagepub.com/content/3/2/149.abstract`.

Joseph Y Halpern. An analysis of first-order logics of probability. *Artificial Intelligence*, 46(3):311–350, 1990. doi: 10.1016/0004-3702(90)90019-V. URL `https://doi.org/10.1016/0004-3702(90)90019-V`.

David Heckerman, Chris Meek, and Daphne Koller. Probabilistic entity-relationship models, PRMs, and plate models. In L Getoor and B Taskar, editors, *Introduction to Statistical Relational Learning*. MIT Press, 2007.

Tamás Horváth, Zoltán Alexin, Tibor Gyimóthy, and Stefan Wrobel. Application of different learning methods to ~~hungarian~~ part-of-speech tagging. In ~~Savso Dvzeroski~~ and Peter Flach, editors, *Inductive Logic Programming: 9th International Workshop, ILP-99 Bled*, pages 128–139. Springer Berlin Heidelberg, Berlin, Heidelberg, 1999.

Tamás Horváth, Stefan Wrobel, and Uta Bohnebeck. Relational Instance-Based Learning with Lists and Terms. *Machine Learning*, 43(1):53–80, 2001. ISSN 1573-0565.

H Khosravi, T Man, J Hu, E Gao, Richard Mar, and O Schulte. Factorbase code. [Online]. Available: URL = `https://github.com/sfu-ml-lab/FactorBase`, 2019.

Tushar Khot, Sriraam Natarajan, and Jude W Shavlik. Relational one-class classification: A non-parametric approach. In *Proceedings AAAI*, pages 2453–2459, Quebec City, Quebec, Canada, 2014. URL `http://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8578`.

Angelika Kimmig, Lilyana Mihalkova, and Lise Getoor. Lifted graphical models: a survey. *Machine Learning*, pages 1–45, 2014.

Mathias Kirsten, Stefan Wrobel, and Tamás Horváth. Distance-based approaches to relational learning and clustering. In ~~Savso Dvzeroski~~ and Nada ~~Lavrave~~, editors, *Relational Data Mining*, pages 213–232. Springer Berlin Heidelberg, 2001.

Arno J Knobbe. *Multi-relational data mining*, volume 145. Ios Press, 2006.

Judice L Y Koh, Mong Li Lee, Wynne Hsu, and Wee Tiong Ang. Correlation-based attribute outlier detection in XML. In *Proceedings ICDE*, pages 1522–1524, Cancun, Mexico, 2008. IEEE. URL `http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=4492792`.

Daphne Koller and Avi Pfeffer. Object-oriented Bayesian networks. In Dan Geiger and Prakash P Shenoy, editors, *Proceedings UAI*, pages 302–313. Morgan Kaufmann, 1997. URL `http://arxiv.org/abs/1302.1554`.

Stefan Kramer, Nada Lavrac, and Peter Flach. Propositionalization approaches to relational data mining. In *Relational Data Mining*, pages 262–286. Springer, 2000. ISBN 3-540-42289-7.

~~Ondvrej Kuvzelka~~ and Filip ~~Zelezny~~. Hifi: Tractable propositionalization through hierarchical feature construction. In *Late Breaking Papers, ILP*, page 69, 2008.

Guiliang Liu and Oliver Schulte. Deep Reinforcement Learning in Ice Hockey for Context-Aware Player Evaluation. In *Proceedings IJCAI*, pages 3442–3448. International Joint Conferences on Artificial Intelligence Organization, 2018. doi: 10.24963/ijcai.2018/478. URL `https://doi.org/10.24963/ijcai.2018/478`.

Joris Maervoet, Celine Vens, Greet Vanden Berghe, Hendrik Blockeel, and Patrick De Causmaecker. Outlier detection in relational data: A case study in geographical information systems. *Expert Systems With Applications*, 39 (5):4718–4728, April 2012. ISSN 0957-4174. doi: 10.1016/j.eswa.2011.09.125. URL `http://dx.doi.org/10.1016/j.eswa.2011.09.125`.

Emmanuel Muller, Ira Assent, Patricia Iglesias, Yvonne Mulle, and Klemens Bohm. Outlier ranking via subspace analysis in multiple views of the data. In *Proceedings ICDM*, pages 529–538, 2012. doi: ICDM.2012.112. URL `https://doi.org/10.1109/ICDM.2012.112`.

Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings IEEE*, 104(1):11–33, 2016. doi: 10.1109/JPROC.2015.2483592.

Frank Nielsen and Richard Nock. On the chi square and higher-order chi distances for approximating f-divergences. *IEEE Signal Processing Letters*, 21(1):10–13, 2014.

Petra Kralj Novak, Nada Lavravc, and Geoffrey I Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *The Journal of Machine Learning Research*, 10:377–403, 2009.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.

Verónika Peralta. Extraction and Integration of MovieLens and IMDb. Technical report, APDM project, 2007.

Matic Perovsek, Anze Vavpetic, Bojan Cestnik, and Nada Lavrac. A wordification approach to relational data mining. In *Proceedings DS*, Lecture Notes in Computer Science, pages 141–154, Singapore, 2013. Springer. doi: 10.1007/978-3-642-40897-7_10. URL `https://doi.org/10.1007/978-3-642-40897-7{\%}5C{\_}10`.

David Poole. First-order probabilistic inference. In *Proceedings IJCAI*, pages 985–991, 2003.

Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. *Proceedings Special Interest Group on Management of Data*, pages 427–438, 2000. doi: 10.1145/342009.335437. URL `https://doi.org/10.1145/342009.335437`.

Fatemeh Riahi and Oliver Schulte. Model-Based Outlier Detection for Object-Relational Data. In *Proceedings SSCI*, pages 1590–1598. IEEE, 2015a. doi: 10.1109/SSCI.2015.224. URL `https://doi.org/10.1109/SSCI.2015.224`.

Fatemeh Riahi and Oliver Schulte. Codes and datasets. [online]. available:. `ftp://ftp.fas.sfu.ca/pub/cs/oschulte/CodesAndDatasets/`, 2015b.

Fatemeh Riahi and Oliver Schulte. Propositionalization for unsupervised outlier detection in multi-relational data. In *Proceedings FLAIRS*, pages 448–453, Key Largo, Florida, USA, 2016. URL `http://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS16/paper/view/12786`.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. Relation extraction with matrix factorization and universal schemas. In *Proceedings HLT-NAACL*, pages 74–84, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA, 2013. URL http://aclweb.org/anthology/N/N13/N13-1008.pdf.

Kurt Routley and Oliver Schulte. A Markov game model for valuing player actions in ice hockey. In *Proceedings UAI*, pages 782–791, 2015.

Sunita Sarawagi, Rakesh Agrawal, and Nimrod Megiddo. Discovery-driven exploration of OLAP data cubes. In *Proceedings EDBT*, pages 168–182, Valencia, Spain, 1998. Springer-Verlag. doi: 10.1007/BFb0100984. URL https://doi.org/10.1007/BFb0100984.

Oliver Schulte. A tractable pseudo-likelihood function for Bayesian networks applied to relational data. In *Proceedings SIAM*, pages 462–473, 2011. doi: 10.1137/1.9781611972818.40. URL https://doi.org/10.1137/1.9781611972818.40.

Oliver Schulte and Sajjad Gholami. Locally consistent Bayesian network scores for multi-relational data. In *Proceedings JCAI*, pages 2693–2700, Melbourne, Australia, 2017. doi: 10.24963/ijcai.2017/375. URL https://doi.org/10.24963/ijcai.2017/375.

Oliver Schulte and Hassan Khosravi. Learning graphical models for relational data via lattice search. *Machine Learning*, 88(3):331–368, 2012.

Oliver Schulte and Kurt Routley. Aggregating predictions vs. aggregating features for relational classification. In *Proceedings CIDM*, pages 121–128, Orlando, FL, USA, 2014. IEEE. doi: 10.1109/CIDM.2014.7008657. URL https://doi.org/10.1109/CIDM.2014.7008657.

Oliver Schulte, Hassan Khosravi, Arthur Kirkpatrick, Tianxiang Gao, and Yuke Zhu. Modelling relational statistics with bayesian networks. *Machine Learning*, 94:105–125, 2014. URL http://link.springer.com/article/10.1007{\%}2Fs10994-013-5362-7.

Tobias Sing, Oliver Sander, Niko Beerenwinkel, and Thomas Lengauer. *ROCR: Visualizing the performance of scoring classifiers.*, 2012. URL http://cran.r-project.org/package=ROCR.

Yizhou Sun, Han Jiawei, and Peixiang Zhao. Rankclus: Integrating clustering with ranking for heterogeneous informationnetwork analysis. In *Proceedings EDBT*, pages 565–576, New York, NY, USA, 2009. ACM.

Guanting Tang, James Bailey, Jian Pei, and Guozhu Dong. Mining multi-dimensional contextual outliers from categorical relational data. In *Proceedings SSDBM*, pages 1171–1192, 2013. doi: 10.3233/IDA-150764. URL https://doi.org/10.3233/IDA-150764.

Stephane Tuffery. *Data Mining and Statistics for Decision Making*. Wiley Series in Computational Statistics, 2011. URL http://ca.wiley.com/WileyCDA/WileyTitle/productCd-0470688297.html.

Daisy Zhe Wang, Eirinaios Michelakis, Minos Garofalakis, and Joseph M Hellerstein. Bayesstore: managing large, uncertain data repositories with probabilistic graphical models. In *Proceedings VLDB*, pages 340–351. VLDB Endowment, 2008. doi: 10.14778/1453856.1453896. URL http://www.vldb.org/pvldb/1/1453896.pdf.

Rongjing Xiang and Jennifer Neville. Relational learning with one network: An asymptotic analysis. In *Proceedings AISTATS*, pages 779–788, 2011.

URL http://proceedings.mlr.press/v15/xiang11a/xiang11a.pdf.