Naming the most Anomalous Cluster in Hilbert Space for Structures with Attribute Information

Janis Kalofolias, Jilles Vreeken

CISPA Helmholtz Center for Information Security
University of Saarland, Saarbrücken, Germany
{janis.kalofolias,vreeken}@cispa.de

Abstract

We consider datasets consisting of arbitrarily structured entities (e.g., molecules, sequences, graphs, etc) whose similarity can be assessed with a reproducing kernel (or a family thereof). These entities are assumed to additionally have a set of named attributes (e.g.: `number_of_atoms`, `stock_price`, etc). These attributes can be used to classify the structured entities in discrete sets (e.g., '`number_of_atoms` < 3', '`stock_price` ≤ 100', etc) and can effectively serve as Boolean predicates. Our goal is to use this side-information to provide named kernel-based anomaly detection. To this end, we propose a method which is able to find among all possible entity subsets that can be described as a conjunction of the available predicates either a) the optimal cluster within the Reproducing Kernel Hilbert Space, or b) the most anomalous subset within the same space. Our method employs combinatorial optimisation of an adaptation of the Maximum-Mean-Discrepancy measure that captures the above intuition. Additionally, we propose a criterion to select the optimal one out of a family of kernels in a way that preserves the available side-information. Finally, we provide several real world datasets that demonstrate the usefulness of our proposed method.

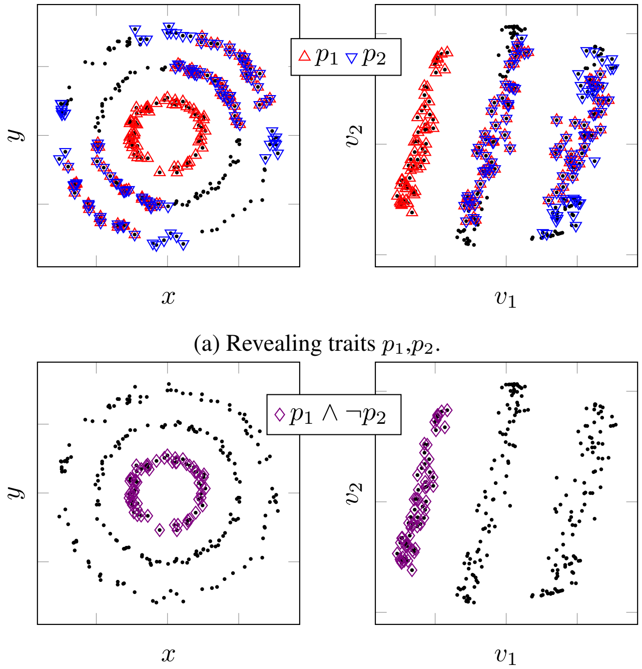
1 Introduction

In the setting we consider we are given a dataset of entities, each of which contains an arbitrary structure that can be represented in a flexible form and consist of multiple dimensions, that can even vary from one entity to the other. For example, proteins, molecules, graphs, images, time series, or effectively any out of the large variety of structures on which a meaningful positive definite kernel can be defined. In addition, we assume that a set of relevant attributes is available for each of the given entities; as such we consider named properties that describe the entities in a way that is sensible and useful to a human user. More specifically, these attributes should (latently, or explicitly) capture interesting traits of the studied entities by which it is possible to meaningfully group them.

Our aim is to find, out of all possible subsets of entities that we can possibly identify by defining conditions on these attributes, that particular subset whose structure deviates as much as possible, either from the rest of the dataset or from the entirety of it. We express these subsets with deviating structure as either a clustering (of 2 clusters) in the given reproducing kernel Hilbert space, or the most anomalous subset within that space. Either way, the result is a named subset of the data that, in addition to the members of the subset itself, is equipped with a meaningful and easily understandable description of this subset.

As an example, consider the active research field of computer-aided drug discovery, where molecules are scrutinised based on their structure and general chemical properties to find potential drug candidates for a given medical condition. A great amount of information on chemical properties for molecules is available, or can be inferred with relative ease via simulations based on the molecule shape, e.g., electronic density, number of atoms, benzene rings, etc; importantly, such chemical and structural properties of molecules can be captured by an appropriate kernel on molecule shapes. However, only a relatively small set of drug-like substances has been meticulously annotated with their drug-related properties by lab specialists. These latter properties could be suggestive traits like toxicity, bioavailability, affinity to a specific target, etc, which can be highly indicative of the fitness of a substance for a medical condition. It would then be of great use to find groups that stand out with respect to their shape-based properties, but that at the same time share a common set of drug-related, revealing traits. In this paradigm, we shift from the discovery of a list of molecules, like paracetamol, ibuprofen, etc, to a set which, in addition, has an easily accessible description based a set of useful properties, like: *'painkillers with high bio-availability and low toxicity'*. Aside from the clear usefulness of this description to a human researcher per se, the molecules that fulfil this description define a landmark in the Hilbert Space, whose proximity to non-annotated data can be assessed with the kernel on molecules.

This paradigm deviates from having one single suggestive trait like toxicity, for which one could fit a classifier on the suggestive traits or the structure, with several methods from supervised learning available to this task. Instead, our goal differs in that there is no singled-out regression or classification trait. We also seek for a simple, concise and exact description, which thus prohibits the use of a linear combination of proper-



(a) Revealing traits p_1,p_2.

(b) Most anomalous named subset: $p_1 \wedge \neg p_2$.

Figure 1: Toy example of points with structure in $\mathbb{R}^2$ captured by a Gaussian kernel, also depicted along the first two eigenvectors $\mathbf{v}_1$, $\mathbf{v}_2$ of the Gramian. Using the two suggestive traits p_1, p_2, we find that the most anomalously structured subset that can be named using these traits is $p_1 \wedge \neg p_2$.

ties: instead of $-0.3 \cdot$ `toxicity_on_rats` $+ 1.2 \cdot$ `tolerance_in_humans`$+0.6 \cdot \ldots$ we seek a description like: $\neg$`toxic` $\wedge$ `tolerance_humans`, which is arguably more accessible to a human than a high-dimensional vector of coefficients. Consider, for example, the toy dataset of entities in Fig. 1a, for which we assume the structural information to be the location of each point in $\mathbb{R}^2$. The arrangement of the points can be captured well by a radial basis function kernel, as seen when we project them on the space spanned by the first two eigenvectors of the kernel (Fig.1b); we also assume the availability of two suggestive traits, `Property 1`, `Property 2`. We now seek to find any exact combination of these properties that results in the named subset that has the most deviating structural properties, where the latter is judged by the kernel. This subset can be named using the logical conjunction $p_1 \wedge \neg p_2$, which is an exact description that is tangible by a human.

It also becomes important to define what those structural properties are that the kernel should deem relevant for its similarity assessment. Although we assume to be given several attributes for the dataset, these cannot directly function as a target variable, like in the typical classification or regression scenaria. A clear cut choice for such a single variable does not apply, since we are not limited to one such parameter. We hence make two key assumptions: that 1) the selection of the properties themselves are axiomatically relevant to the dataset, by the mere fact that they were selected to be included in it, and 2) the similarity of two subsets of entities that arise from a logical combination of these properties is dictated by the similarity of these two combinations. We use these assumptions to quantify the degree of fitness for a kernel that at the same time takes into consideration all available properties as Boolean functions. We then use this measure to select a good fit for a kernel, and study schemes for multiple kernel learning where simpler kernels are linearly combined into a more fit one.

Thus, the main contributions of this paper are as follows.

(a) We propose a family of objective functions on subsets of entities with structure, whose value is high only when the entity structures in the subset stand out—according to a positive definite kernel—from those in the dataset.

(b) We provide an upper bound for our objective functions and use it in an adapted branch-and-bound solver for their efficient and exact combinatorial optimisation.

(c) We provide methods to tune the hyper-parameter of the involved kernel, which takes into account all available properties of the dataset at the same time.

Note that due to lack of space, we delegate all proofs of Sec. 3 to the available[1] extended version of this work.

2 Preliminaries

In this work we study datasets that consist of a set of entities $E = \{e_1, \ldots, e_n\}$, for each of which a set of attributes is defined. Each entity is accompanied with such additional information that can be used to evaluate a similarity between these entities through a positive definite kernel $\kappa : E \times E \to \mathbb{R}$. The attributes themselves are accessible to our algorithm through a set of predicates $P = \{p_1, \ldots, p_\mu\}$ which are derived from the attributes and thus have an interpretable description. We choose to split the numerical values into 5 ranges (`very_low`, `low`, `normal`, `high`, `very_high`), so that each interval contains roughly the same number of entities. On categorical and Boolean attributes we use the one-hot encoding. For instance, we will later study a dataset that consists of a set of traded stocks as entities, for each of which there are available financial attributes (e.g., payed dividends, time listed, industry sector, etc.); these we discretise to form predicates like [`pays_dividends`], [`time_listed` $<$ 1 `year`], [`sector` $=$ `pharmaceutics`].

Formally, the predicates are Boolean functions $p_j : E \to \{\top, \bot\}$, for $\top$, $\bot$ the `true` and `false` conditions, respectively. The extension operation $\operatorname{ext} : P \to 2^E$ uses a predicate as a set characteristic function over E and thereby partitions the entities E into the set of those which satisfy the predicate $\operatorname{ext}(p) := \{e \in E \mid p_j(e) = \top\}$, and those which do not $E \setminus \operatorname{ext}(p)$. These named predicates can be further combined into conjunctions, which form more detailed descriptions; we refer to such a description as a selector s, and express it as the subset of the involved predicates $s \subseteq P$. Naturally, the cardinality $|s|$ of a selector is the number of predicates in its description. We also use the generalisation of the extension operation on selectors $\operatorname{ext}(s) := \bigcap_{p \in s} \operatorname{ext}(p)$, and define the related operator of

[1] Accessible at https://eda.mmci.uni-saarland.de/prj/nuts.

intention $\mathrm{int} : E \to 2^P$ which assigns to each entity all predicates that it satisfies $\mathrm{int}(e) := \{p \in P \mid e \in \mathrm{ext}(p)\}$.

In our example, each of the stocks comes with additional structure: a time series of its recent daily prices. Among among all combinations of available financial attributes, we now seek the one that describes the subset of stocks with the most deviating daily price progression from that of the entire dataset. We measure this deviation by a state-of-the-art positive definite kernel on time-series. Hence, we first define an appropriate kernel-aware measure of anomaly for entity subsets, and then optimise it over the language of named entity subsets $\mathcal{L} := \{\mathrm{ext}(s) \mid s \subseteq P\}$.

3 Most Outstanding Named Entity Subset

We now formalise our intuition into an optimisation problem of a kernel-aware measure of anomaly for entity subsets. We hence focus on the distribution of the structural information of the entities; we search for the particular subset $Q \subseteq E$ with the maximally deviating such distribution compared to either 1) the same distribution over the entire E or 2) within only the complement $\bar{Q}$. The first requirement is compatible with the assumption that our dataset comprises the entire population; therefore any quantity computed on E is deterministic and not affected by the choice of Q. On the contrary, if we assume E to be just a sample of the entire population, then to avoid unnecessary bias we may only compare statistics computed on independent samples; since one is Q, the other must only contain the entities $E \setminus Q$. We hence refer to the first problem as anomalous discovery and to the latter as contrastive. The next step is to specify how to employ a positive definite kernel to measure the distributional distance between the corresponding two sets.

Maximum Mean Discrepancy (Gretton et al. 2007) Exactly for this task Gretton et al. (2007) provide a special case for a known result from real analysis (Dudley 2002, Lemma 9.3.2). This informally states that for any two (Borell) probability measures p, q, defined over a metric space $\mathcal{X}$, the following holds. Consider a dense enough space of transformations $g : \mathcal{X} \to \mathbb{R}$, like the set of all bounded functions over $\mathcal{X}$; then it is $p = q$ if and only if the expectation of every transformation under p and under q coincides $\mathbb{E}_{x\sim p}[g(x)] = \mathbb{E}_{x\sim q}[g(x)]$. In the role of a dense enough space, Gretton et al. (2007) propose to use the unit ball in the Hilbert Space of a reproducing kernel. The resulting measure between two distributions p, q is the Maximum Mean Discrepancy (MMD)

$$\mathrm{MMD}(p,q) := \|\mu(p) - \mu(q)\|_{\mathcal{H}}\,,$$
$$\mu(\cdot) := \mathbb{E}_{x\sim\cdot}[\phi(x)]\,, \qquad \hat{\mu}(P) = \frac{1}{|P|}\sum_{x\in P}\phi(x) \qquad (1)$$

where $\phi : \mathcal{X} \to \mathcal{H}$ is the feature map of a given kernel on $\mathcal{X}$ and μ is the mean of all points in $\mathcal{H}$ to which ϕ maps the elements of $\mathcal{X}$. An empirical mean $\hat{\mu}$ can also be estimated over a finite subset $P \subset \mathcal{X}$ that was itself sampled under p. This gives rise to the more relevant (squared) biased empirical estimator of MMD, which for two sets Q, Q' is

$$\widehat{\mathrm{MMD}}^2(Q,Q') = \frac{1}{|Q|^2}\sum_{e,e'\in Q}\kappa(e,e') - \frac{2}{|Q||Q'|}\sum_{\substack{e\in Q,\\ e'\in Q'}}\kappa(e,e') + \frac{1}{|Q'|^2}\sum_{e,e'\in Q'}\kappa(e,e')\,, \qquad (2)$$

where $\kappa(e,e') = \langle\phi(e),\phi(e')\rangle_{\mathcal{H}}$. We hence adopt this measure to quantify the dissimilarity between Q and its pair.

An objective for our task In contrast, however, to the setting for which the MMD was developed, we need to evaluate as candidate sets $Q \in \mathcal{L}$ all those that result from a named combination of the dataset attributes. This means that without proper scaling, selecting just an outlier could trigger a false discovery. We therefore adapt the $\widehat{\mathrm{MMD}}$ by multiplying it with a scaling factor $a(|Q|)$, which depends only on the size of Q and can be interpreted as a size prior. This yields our objective function

$$f_t(Q;\kappa,\gamma) := a_t^{\gamma}(|Q|)\cdot\widehat{\mathrm{MMD}}^2_{\kappa}(Q,Q'_t)\,, \quad \gamma > 0\,, \qquad (3)$$

where $t = \text{ano}$ and $t = \text{con}$ indicate the anomalous and contrastive assumptions, respectively, for which we define

$$a_{\text{ano}}(m) := m \qquad Q'_{\text{ano}} := \quad E$$
$$a_{\text{con}}(m) := \frac{m(n-m)}{n} \qquad Q'_{\text{con}} := E \setminus Q\,.$$

The scalar γ is a tuning parameter that controls the relative importance between the prior on the subset cardinality and the deviation component of the objective.

Note that we are not limited to using the given priors and any reasonable choice will do. The intuition of our own choice is that larger sets are less prone to be outliers and are generally more informative; alternatively, from a statistical perspective, larger sets should lead to stricter confidence bounds of the $\widehat{\mathrm{MMD}}$ statistic. In the contrastive case, due to the symmetry $f_{\text{con}}(Q) = f_{\text{con}}(\bar{Q})$ we wish for $|Q|$ to be far from both extremes. When $t = \text{ano}$ our simpler choice suffices due to the self-limiting effect of increasing the size of Q, during which its distribution grows closer to that of E. Importantly, in this case (and for γ=1) we can write

$$\sqrt{f_{\text{ano}}(Q;\kappa,\gamma=1)} \propto \frac{\|\hat{\mu}(Q) - \mu(E)\|_{\mathcal{H}}}{\sigma/\sqrt{|Q|}}\,,$$

where $\hat{\mu}$ is as defined in Eq. (1) and σ^2 is the variance of ϕ. Since we assumed that E is the full population, $\mu(E)$ is the true mean and our score is proportional to the (squared) z-score of the empirical mean estimator $\hat{\mu}(Q)$, after we generalise the absolute difference to be the norm in $\mathcal{H}$.

We now reformulate our objective to reveal structure that is convenient for what follows and to further show that both problems differ only in the set cardinality prior $a_t(|Q|)$.

Lemma 3.1. *Let $m_Q := |Q|$ be the cardinality of any entity subset. Then we can write our objective of Eq.* (3) *as*

$$f_t(Q;\kappa,\gamma) = a_t^{\gamma-2}(m_Q)\mathbf{z}_Q{}^{\top}K\mathbf{z}_Q\,, \qquad (4)$$

where $\mathbf{K} \in \mathbb{R}^{n \times n}$ is the Gramian $\mathbf{K}_{i,j} := \kappa(e_i, e_j)$ and

$$\mathbf{z}_Q := \mathbf{c}_Q - \frac{m_Q}{n}\mathbf{e}\,,$$

for $\mathbf{e} := (1, \ldots, 1) \in \mathbb{R}^n$ the vector of all ones and $\mathbf{c}_Q := (\mathbb{1}\{e_i \in Q\})_{i=1}^n$ the characteristic[2] vector of set Q.

We can now formalise our problem as follows.

Problem 1. *Given dataset E with attributes yielding the predicates P and with structure captured by kernel κ, solve*

$$\max_{Q \in \mathcal{L}} f_t(Q; \kappa, \gamma)\,.$$

This is a hard combinatorial problem and can be solved optimally by a branch-and-bound algorithm that we adapt to traverse the set $\mathcal{L}$, described in the online appendix and openly implemented[3]. We next focus on an appropriate upper bound that is necessary to efficiently use this algorithm.

An Upper Bound for our Objective

We now derive an upper bound for Eq. (4) that can be computed in linear time, assuming an one-time sorting operation with a time complexity of $O(n \log n)$.

Formally, we seek a function $\hat{f} : 2^E \to \mathbb{R}$ that when evaluated at an entity subset $Q \subseteq E$ computes an upper bound of the objective over all subsets of its argument, $\hat{f}(Q) \geq \max_{R \subseteq Q} f(R)$. Such a bound can be computed in two steps: first we can bound the objective exclusively over all subsets $R \subseteq Q$ with a fixed cardinality m_R, and then we can compute an upper bound for all subsets as the maximum of all cardinality-constrained maxima. We hence seek an upper bound of the sub-problem

$$\hat{f}_t(Q; \kappa, \gamma, m) \geq \max_{R \subseteq Q\,,\, |R| = m} f_t(R; \kappa, \gamma)\,. \tag{5}$$

Since now the size m_Q remains constant, we can derive a bound for each sub-problem as follows. Let $\mathbf{e}_i$ denote the i-th vector of the standard basis, i.e., the vector with a single one at the i-th position, and define $\mathbf{e}_{:m} := \sum_{i=1}^{m} \mathbf{e}_i$. Let $\mathbf{v}_1, \ldots, \mathbf{v}_k$ be the eigenvectors of $\mathbf{K}$ with rank $k \leq n$ and corresponding eigenvalues $\lambda_1 \geq \ldots, \lambda_k$. Further, denote $\mathbf{v}_{i\uparrow[Q]}$, $\mathbf{v}_{i\downarrow[Q]}$ the vector with those entries of $\mathbf{v}_i$ for which the characteristic function of $\mathbf{c}_Q$ is non-zero, sorted in increasing and decreasing order, respectively.

Lemma 3.2. *Given any integer constant $\rho < k$, an upper bound for the problem in Eq.* (5) *is*

$$\hat{f}_t(Q; \kappa, \gamma, m) =$$

$$a_t^{\gamma-2}(m_Q) a_{con}(m_Q) \left(\sum_{i=1}^{\rho} \lambda_i \min\{u_i, \vec{u}_i\} + \lambda_{\rho+1} \vec{u}_{\rho+1} \right)\,,$$

where $\vec{u}_i := \max\left\{0, 1 - \sum_{j=1}^{i-1} u_j\right\}$ and

$$u_i := \frac{\left(\max\left\{\mathbf{e}_{:m}{}^{\top}\mathbf{v}_{i\uparrow[Q]}, \mathbf{e}_{:m}{}^{\top}\mathbf{v}_{i\downarrow[Q]}\right\} - \frac{m}{n}\mathbf{e}^{\top}\mathbf{v}_i\right)^2}{a_{con}(m)}\,.$$

We can now compute a bound $\hat{f}_t(Q; \kappa, \gamma)$ over all subsets $R \subseteq Q$ using Lemma 3.2 as follows.

$$\begin{aligned} \hat{f}_t(Q; \kappa, \gamma) &= \max_{m \in \{0, \ldots, m_Q\}} \hat{f}_t(Q; \kappa, \gamma, m) \\ &\geq \max_{m \in \{0, \ldots, m_Q\}} \max_{R \subseteq Q,\, |R| = m} f_t(R) \geq f_t(Q)\,. \end{aligned}$$

In the supplementary material we provide an algorithm that uses this approach to compute an upper bound in $O(n\rho)$ time, which is linear when ρ is considered a small constant.

4 Hyperparameter Optimisation

Next, we present methods to tune the hyper-parameters of the kernel that assesses the entity similarity, in a way that preserves important information for the task of anomaly detection or clustering, as introduced in Sec. 3.

Since, however, there is no explicitly defined target variable available for classification or regression, it is not possible to use standard schemes from supervised learning. Instead, we are given a set of predicates P, each of which can be seen as a classification variable. We hence make two key assumptions: 1) *the attributes of the datasets and thereby the predicates derived from them are relevant to the task for which the dataset was created*, and 2) *two subsets whose predicate description is similar should themselves be similar*. We therefore seek a method which takes into consideration all predicates at the same time without favouring just a single one or a few of them, and that admits a meaningful interpretation of predicate conjunctions.

Assessing the fitness of a candidate kernel on E

We now consolidate these assumptions into a method to assess the fitness of a candidate kernel $\kappa : E \times E \to \mathbb{R}$.

The first obstacle is that we require to evaluate the performance of a kernel over entities, using ground truth over predicates. An straightforward way to induce a similarity over entities using their predicates is by using the intentions of each entity as a feature.

$$\kappa_{\text{lp}}(e_1, e_2) := \frac{|\operatorname{int}(e_1) \cap \operatorname{int}(e_2)|}{\sqrt{|\operatorname{int}(e_1)||\operatorname{int}(e_2)|}}\,,$$

This is equivalent to the normalised intersection kernel, or the linear kernel over the characteristic vectors of the entity intentions. In addition to the downsides of the linear kernel, this kernel loses discriminating power when the number of predicates is small, and increases the complexity of the method to that of $|E|$, which is typically larger than $|P|$.

Instead, our method relies on three key steps: first, we capture the similarity of each predicate to the others using a Tanimoto kernel (Tanimoto 1958), which operates on the predicate set P. Then, we 2) use a kernel over sets to compute the candidate kernel, whose domain is the set of entities E, to define one over P. Finally, 3) we assess the fitness of the candidate kernel, as the alignment of the previously derived kernel and the Tanimoto kernel. This process yields a fitness value for any candidate kernel on E, which can subsequently be used as a proxy by a method to either select one kernel out a family thereof, or to create a composite kernel.

[2] Here, $\mathbb{1}\{\cdot\} = 1$ if the condition $\cdot$ is satisfied and 0 otherwise.

[3] Available at https://eda.mmci.uni-saarland.de/prj/nuts.

The Tanimoto kernel (Tanimoto 1958) is the usual name of the Jaccard Index when used as a kernel, and in its original form operates on the power-set of a ground set $\kappa_{\text{Tan}} : 2^E \times 2^E \to [0,1]$, here on the space consisting of all entity subsets. We can therefore apply it on the set of predicates P through the use of the extension operator

$$\kappa_{\text{Tan}}(p_1, p_2) := \frac{|\operatorname{ext}(p_1) \cap \operatorname{ext}(p_2)|}{|\operatorname{ext}(p_1) \cup \operatorname{ext}(p_2)|}, \quad p_1, p_2 \in P .$$

The Tanimoto kernel is known to be positive definite (Gower 1971) and captures the normalised amount of shared structure between sets. This makes κ_{Tan} a natural choice for predicate conjunctions, and can therefore measure the similarity over P. Owing to the assumption of meaningful selection of predicates, we now treat their similarity as ground truth.

The candidate kernel κ operates on two entities. In order to compare it with the ground truth we need to transform it into a kernel that operates on sets of entities. Arguably the most appropriate choice is the kernel mean map (Muandet et al. 2017) of κ. This is in turn also a positive definite kernel, and is based on the same assumptions used in Sec. 3 to derive the MMD: that each set contains i.i.d. samples of a distribution, which is mapped in the Hilbert Space to the mean of the mappings of each element in the set. This is exactly the kernel that induces the MMD distance[4], and therefore fits naturally to our assumptions. When this kernel is applied on any two predicates $p_1, p_2 \in P$ it becomes

$$\kappa_{\text{mm}}(p_1, p_2; \kappa) = \frac{1}{|\operatorname{ext}(p_1)||\operatorname{ext}(p_2)|} \sum_{\substack{e_1 \in \operatorname{ext}(p_1) \\ e_2 \in \operatorname{ext}(p_2)}} \kappa(e_1, e_2) .$$

We now need to compare the ground truth similarity over P to the one induced by the candidate kernel κ through the kernel mean map embedding, again over P. For this we use an established similarity measure of two kernels, the kernel alignment (Baumgartner, Böhm, and Baumgartner 2005)

$$\operatorname{algn}(K_1, K_2) := \frac{\langle K_1, K_2 \rangle_{\text{F}}}{\sqrt{\|K_1\|_{\text{F}} \cdot \|K_2\|_{\text{F}}}} ,$$

where K_1, K_2 are the Gramians of the two kernels and $\langle \cdot, \cdot \rangle_{\text{F}}$ the Frobenius inner product[5]. This measure resembles the cosine similarity of the two Gramians with respect to the Frobenius inner product, and it is $0 \leq \operatorname{algn}(K_1, K_2) \leq 1$, where the lower bound is due to the definiteness of the Gramians. The upper arises from the Cauchy-Schwarz inequality, and therefore $\operatorname{algn}(K_1, K_2) = 1 \iff K_1 \propto K_2$.

Combining all the above, we define the kernel fitness of κ

$$\operatorname{kernfit}(\kappa; P) := \operatorname{algn}\Big(\mathbf{K}_{\text{Tan}}(P), \mathbf{K}_{\text{mm}}(P; \kappa)\Big) , \qquad (6)$$

where $P \subset 2^E$ is a set of predicates, $\mathbf{K}_{\text{Tan}}(P)$ is the Gramian of the Tanimoto kernel over P, and $\mathbf{K}_{\text{mm}}(P, \kappa)$ is the Gramian of the kernel mean map $\kappa_{\text{mm}}(p_1, p_2; \kappa)$ for each $p_1, p_2 \in P$ and with κ the candidate kernel over entities.

We can now use the kernel fitness defined in Eq. (6) to select the best out of a family of kernels on E, for instance by an appropriate global optimisation scheme, such as grid-search or—as in our experiments—Bayesian optimisation.

Multiple Kernel Learning

A special case of measuring kernel fitness arises when the family of kernels we evaluate is a (positive) linear combination of a collection of constituent kernels. In this case, there exists a non-negative vector of coefficients $\alpha \in \mathbb{R}_+^p$ such that the candidate kernel can be written as $\kappa_\alpha := \sum_{\iota=1}^p \alpha_\iota \kappa_\iota$; then the (squared) kernel fitness of Eq. (6) becomes

$$\operatorname{kernfit}^2(\kappa_\alpha) = \frac{1}{\|\mathbf{K}_{\text{Tan}}\|_{\text{F}}^2} \frac{\alpha^\top \mathbf{v} \mathbf{v}^\top \alpha}{\alpha^\top \mathbf{W} \alpha} ,$$

where $\mathbf{W} \in \mathbb{R}^{p \times p}$ with $\mathbf{W} \succeq 0$ and $\mathbf{v} \in \mathbb{R}^p$, defined as

$$\begin{aligned} W_{i,j} &:= \langle \mathbf{K}_{\text{mm}}(\kappa_i), \mathbf{K}_{\text{mm}}(\kappa_j) \rangle_{\text{F}} \\ v_i &:= \langle \mathbf{K}_{\text{mm}}(\kappa_i), \mathbf{K}_{\text{Tan}} \rangle_{\text{F}} \end{aligned} , \quad i, j = 1, \ldots, p .$$

When the components κ_ι are guaranteed to be orthogonal Cristianini et al. (2002) provides an optimal solution for α, which amounts to using a vector of coefficients whose elements are proportional to the alignment of each component. In practice, however, our candidate kernels can be not only non-orthogonal, but also highly correlated, which makes the $\mathbf{W}$ matrix badly conditioned or even non-invertible. For these cases we modify the the solution that is optimal in the orthogonal case, in what forms the following heuristic.

We keep adding the components, considering them in decreasing order of their alignment. At each step, the added components are weighted with coefficients which are proportional to their fitness, which can be shown to be $\operatorname{kernfit}(\kappa_\iota) \propto {}^{(\alpha_\iota v_\iota)^2}/_{W_{\iota,\iota}}$. In the end we pick the topmost coefficients from the beginning until the index that maximises the kernel fitness. Although, when orthogonal components are added the resulting alignment can only increase (Cristianini et al. 2002), adding a component that is correlated with an already added one may lower the resulting fitness. Thus, our selection scheme results in a sparse selection in case of highly correlated components, while it remains optimal in case of orthogonal ones.

5 Related Work

From the perspective of the general goal, our method can be seen as a special case of Subgroup discovery (Herrera et al. 2011). In this setting, a dataset with tabular attributes is given alongside a clearly defined target variable. This variable is then maximised using a fitting objective function. This setting deviates from ours in that we do not have such a target variable. Further, our method can incorporate a widely used family of weighted accuracy/impact objectives (Atzmueller 2015) by using as a special case the linear kernel.

Perhaps more relevant is the special case of exceptional model mining (Duivesteijn, Feelders, and Knobbe 2016), which uses p-values or information-theoretic measures to assess whether the parameters of two models trained on these target variables are the same: one trained on the subset and

[4] The distance induced by a kernel is $d(p_1, p_2) := \kappa(p_1, p_1) + \kappa(p_2, p_2) - 2\kappa(p_1, p_2)$, which can be verified to match Eq. (2).

[5] It is $\langle K_1, K_2 \rangle_{\text{F}} = \operatorname{Tr}\big[{K_1}^\top K_2\big]$ and $\|K\|_{\text{F}}^2 = \sum_{i,j} K_{ij}^2$.

the other on the entire dataset. These models either require a well defined relation between the target variables (e.g., measuring correlation, using tests, etc), or one target must function as a classification label, so that a linear model can be trained over the rest of the variables. These methods, however, do not admit an optimistic estimate for efficiently exact optimisation, while they have not studied the use of kernels.

From the optimisation perspective, several methods have been used, which can be seen as variations of the Branch and Bound algorithm, or greedy variants like beam search. We value an exact maximisation with concrete bounds, and so we focus in the former family of methods, which require optimistic bounds. The task of computing the optimistic estimator of Eq. (5) is an integer optimisation problem for a fraction of quadratic functions. Due to its resemblance with the Rayleigh quotient, our problem also becomes relevant to maximisation and minimisation schemes, by simply inverting the fraction. Several methods can solve the unconstrained problem (Konno 1980; Li, Sun, and Liu 2012), however the cardinality constraint makes them non-applicable. Note that, contrary to continuous optimisation, it is not easy to first solve for the transformation z and then solve for the x in the unit box. Indeed, these methods rely on the particular structure of the 0-1 box, which is violated by the transformation we require. Note that exactly because of the arbitrary scaling function a, generic bounds are not applicable, even the known naïve ones (Shor 1987). A simple method could sort the values of the matrix, but our proposed bound is tighter and more computationally efficient, which makes a comparison equivalent to creating a strawman to later defeat.

When it comes to hyper-parameter optimisation, several methods have been proposed for un-supervised tuning parameters, which can be used for kernel clustering (Langone et al. 2016), or for general clustering (Meila 2018). These methods, however, often require multiple clusters instead of just two, and also ignore the predicate information.

6 Experiments

We implement and evaluate our method on real world datasets. Here we demonstrate our results.

Datasets Despite the abundance of structured datasets (e.g., containing images, graphs, time-series, etc) and similarly many with tabular data, there is a substantial scarcity of datasets with both such information at the same time. We thus compile three datasets that come close to practical tasks from drug discovery, finance, and social sciences, and demonstrate related aspects from our hyper-parameter optimisation methods in Sec. 4. We next quickly outline the nature of these datasets. Detailed parameters and complete results are delegated to the extended version of this work.

In `Chem` we refer to a dataset of drug-like molecules from the *ChEMBL* (Gaulton et al. 2012) database, i.e., substances with potential pharmaceutical usage; their predicates are derived from suggestive pharmaceutical traits (Malone et al. 2010) annotated by human specialists, and their structure is assessed with a pre-computed kernel (Cincilla, Thormann, and Pons 2010). In `Stock` we describe stocks of companies listed in the New York Stack Exchange with indicative financial traits of each company, alongside a time-series of daily prices; these are used to assess stock similarity through extracted Rocket features (Dempster, Petitjean, and Webb 2020). Finally, `Twitter` contains Twitter ego nets (Leskovec and Krevl 2014): small subgraphs of the interaction network centered around selected individuals. Their attributes are followed users and used hash-tags. We compare their graphs using the state-of-the-art Wasserstein-Weisfeiler-Lehman kernel (Togninalli et al. 2019).

Kernel Hyperparameter Tuning Except for the `Chem` dataset, which comes with a pre-computed kernel provided by the PubChem interface, we need to specify hyperparameters for the kernels of our datasets. We therefore demonstrate here the methods introduced in Sec. 4, for both settings of single parameter and multiple kernel learning.

For the `Twitter` dataset we use the state-of-the-art WWL (Togninalli et al. 2019) kernel, κ_{wwl}, which requires the specification of a single scalar parameter γ_{wwl}. We choose this parameter by optimising the kernel fitness of Eq. (6) using Bayesian optimisation with a Gaussian process prior (Snoek, Larochelle, and Adams 2012) evaluated at 120 points (Fig. 3). For `Stock` each of the 1000 extracted Rocket features (Gaulton et al. 2012) yields a radial basis kernel whose σ parameters are individually tuned with the same procedure, resulting in a collection of equally many candidates for multiple kernel learning, which are highly correlated. To then combine these features into a single kernel we use the algorithm described in Sec. 4, which results in a sparse combination of only 4 sub-kernels (Fig. 4a).

As a measure of fitness for this method we also compare the average recall of the classification of each predicate as a classification variable, using a kernel with a parameter trained at each point. We show that the kernels chosen with $\mathrm{kernfit}$ yield significantly higher scores than when maximising the alignment of the linear predicate kernel in Fig. 3b for a broad range of γ_{wwl} values, and for the optimal multiple kernel coefficients arising from the two methods in Fig 4b.

Necessity of constrained optimisation We further motivate our method by demonstrating the necessity of constrained optimisation. Since unconstrained clustering is description-unaware it is extremely unlikely to yield a describable subset in the first place. Finding for it the closest description may also result in a low-quality subset. In Fig. 5 we show the centroids of the optimal named subset Q, a local optimum Q_{km} found by kernel k-means initialised with Q, and the closest named Q_{Jac} in the Jaccard sense to Q_{km}. Unsurprisingly Q_{km} scores the highest in our objective but has no description, while the naïvely found named one has 5 times lower quality than our optimum.

Efficiency of computation To optimise Problem 1 we use our branch and bound variant, for which a key factor deciding its efficiency is the ability of the optimistic estimator to prune the search space. That means that a key measure of this efficiency for a given dataset is the number of states it visits. Since there are no baseline optimistic estimators for our novel objective, as a comparable measurement we show (Fig. 6) the percentage of visited states over those

γ	Subset Description	$\lvert Q\rvert$	mmd
(0,0.5]	$[49 \leq \texttt{price}] \wedge [\texttt{sector} = \texttt{Energy}] \wedge [10 \leq \texttt{MarkCap}]$	0.005	0.2767
(0.5,0.6]	$[1.9 \leq \texttt{lastDiv}] \wedge [\texttt{sector} = \texttt{Energy}] \wedge [9.8 \leq \texttt{MarkCap}]$	0.008	0.2315
(0.6,0.9]	$[\texttt{sector} = \texttt{Energy}] \wedge [\texttt{activeTrading}] \wedge [4.8 \leq \texttt{AvgVol}]$	0.074	0.0548
(0.9,1]	$[\texttt{sector} = \texttt{Energy}] \wedge [\texttt{activeTrading}]$	0.082	0.05
(1,1.25]	$[10 \leq \texttt{price}] \wedge [0.52 \leq \texttt{beta}] \wedge [0.00017 \leq \texttt{lastDiv}]$	0.529	0.0064
(1.25,3.6)	$[10 \leq \texttt{price}] \wedge [0.52 \leq \texttt{beta}]$	0.691	0.0045
(3.6,8]	$[10 \leq \texttt{price}]$	0.834	0.0023

Table 1: Selected subsets from Stock and their metrics.

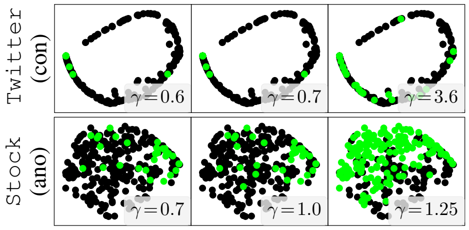
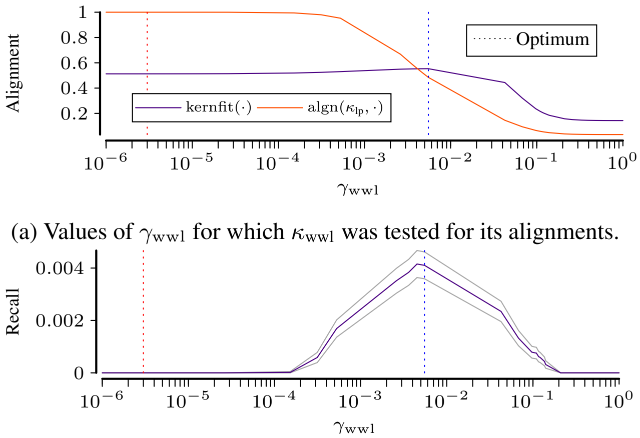


Figure 2: Selected discovered named subsets.



(a) Values of γ_{wwl} for which κ_{wwl} was tested for its alignments.

(b) Classification recall of predicates using $\kappa_{\text{wwl}}(\cdot)$ in a SVM.

Figure 3: Tuning the scalar parameter γ_{wwl} for Twitter, tested at 100 points selected through Bayesian optimisation [above], and the recall of predicate validities using the kernel for the given γ_{wwl}. Average of 50 splits per predicate.[below].

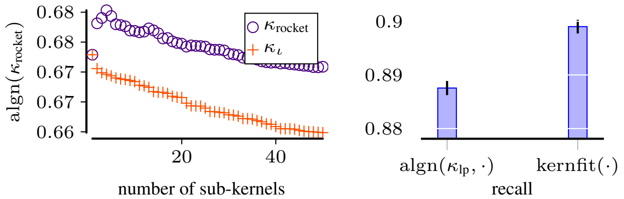
required by an exhaustive search. Our optimistic estimator remains practically efficient even for higher rank matrices. Also, since we can only expressed our objective as an integer quadratic problem (IQP) for a fixed cardinality (see Eq. (5)), the full problem would require solving $O(n)$ hard IQP sub-problems. As we show in the extended version of this work, our method is superior to the IQP approach.

Named subsets A selection of discovered subgroups are listed in Table 1; selected subsets are shown (along the first 2 eigenvectors) in Fig. 2. Quantitatively, the size of the discovered subsets is controlled by the tuning parameter γ, with a simultaneous lowering of the measure of dissimilarity in $\mathcal{H}$.

Of special interest is the occurrence of the predicate $[\texttt{sector} = \texttt{energy}]$ in several top subsets, indicating that the stock prices in this sector were the most deviating from the rest. Since we chose the price sequence in the data to cover the years of the pandemic, they highly overlap the period of heavy restrictions imposed in transport, which has been extensively reported to financially impede this sector.

7 Summary and Discussion

We provide all key components for a method that is able to find named anomalous subsets from datasets of entities that have not only attributes but also such structural information



(a) Optimal MKL coefficients.

(b) Classification score.

Figure 4: Multiple kernel learning for Stock: top-ranking sub-kernels are added until the resulting alignment stops increasing [left], and classification recall of the optimal kernel, compared against the naïve linear predicate kernel κ_{lp}.

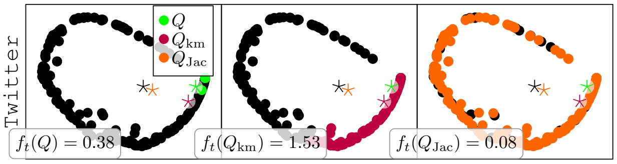


Figure 5: Naming the clusters of kernel k-means yields low quality subsets. Comparison of our discovered, optimal subset Q [left], the kmeans discovered subset without name Q_{km} [middle], and the closest named to the latter Q_{Jac} [right].

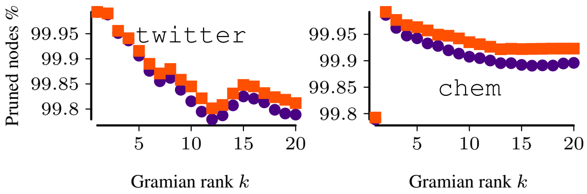


Figure 6: Efficiency of the optimisation in terms of visited search states during the branch and bound algorithm, as the matrix rank k of the Gramian increases.

that can be captured by a reproducing kernel. We motivate two concrete formulations of our problem which stem from reasonable statistical assumptions; we justify the selection of the priors in each formulation and study their parameters.

We show how to practically and efficiently tune the hyper-parameters respecting the assumptions of our task and demonstrate them in real world datasets. Finally, we show our method to be practical and give meaningful results.

References

Atzmueller, M. 2015. Subgroup Discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 35–49.

Baumgartner, C.; Böhm, C.; and Baumgartner, D. 2005. Modelling of classification rules on metabolic patterns including machine learning and expert knowledge. *Biomed. Inf.*, 38(2): 89–98.

Cincilla, G.; Thormann, M.; and Pons, M. 2010. Structuring Chemical Space: Similarity-Based Characterization of the PubChem Database. *Molecular Informatics*, 37–49.

Cristianini, N.; Shawe-Taylor, J.; Elisseeff, A.; and Kandola, J. 2002. On Kernel-Target Alignment. In *Advances in Neural Information Processing Systems*. MIT Press.

Dempster, A.; Petitjean, F.; and Webb, G. I. 2020. ROCKET: Exceptionally Fast and Accurate Time Series Classification Using Random Convolutional Kernels. *Data Mining and Knowledge Discovery*, 1454–1495.

Dudley, R. M. 2002. *Real Analysis and Probability*. Cambridge University Press, second edition.

Duivesteijn, W.; Feelders, A. J.; and Knobbe, A. 2016. Exceptional Model Mining: Supervised Descriptive Local Pattern Mining with Complex Target Concepts. *Data Mining and Knowledge Discovery*, 47–98.

Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; and Overington, J. P. 2012. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Research*, D1100–1107.

Gower, J. C. 1971. A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, 857–871.

Gretton, A.; Borgwardt, K. M.; Rasch, M.; Schölkopf, B.; and Smola, A. J. 2007. A Kernel Method for the Two-Sample-Problem. In *Advances in Neural Information Processing Systems*, 513–520.

Herrera, F.; Carmona, C. J.; González, P.; and del Jesus, M. J. 2011. An Overview on Subgroup Discovery: Foundations and Applications. *Knowledge and Information Systems*, 495–525.

Konno, H. 1980. Maximizing a Convex Quadratic Function Over a Hypercube. *Journal of the Operations Research Society of Japan*, 171–189.

Langone, R.; Mall, R.; Alzate, C.; and Suykens, J. A. K. 2016. Kernel Spectral Clustering and Applications. In *Unsupervised Learning Algorithms*, 135–161. Springer International Publishing. ISBN 978-3-319-24211-8.

Leskovec, J.; and Krevl, A. 2014. *SNAP Datasets: Stanford Large Network Dataset Collection*.

Li, D.; Sun, X. L.; and Liu, C. L. 2012. An Exact Solution Method for Unconstrained Quadratic 0–1 Programming: A Geometric Approach. *Journal of Global Optimization*, 797–829.

Malone, J.; Holloway, E.; Adamusiak, T.; Kapushesky, M.; Zheng, J.; Kolesnikov, N.; Zhukova, A.; Brazma, A.; and Parkinson, H. 2010. Modeling Sample Variables with an Experimental Factor Ontology. *Bioinformatics*, 1112–1118.

Meila, M. 2018. How to Tell When a Clustering Is (Approximately) Correct Using Convex Relaxations. *Advances in Neural Information Processing Systems*.

Muandet, K.; Fukumizu, K.; Sriperumbudur, B.; and Schölkopf, B. 2017. Kernel Mean Embedding of Distributions: A Review and Beyond. *Foundations and Trends® in Machine Learning*, 1–141.

Shor, N. Z. 1987. Class of Global Minimum Bounds of Polynomial Functions. *Cybernetics*, 731–734.

Snoek, J.; Larochelle, H.; and Adams, R. P. 2012. Practical Bayesian Optimization of Machine Learning Algorithms. *Advances in Neural Information Processing Systems 25*, 2951–2959.

Tanimoto, T. T. 1958. *An Elementary Mathematical Theory of Classification and Prediction by T.T. Tanimoto*. International Business Machines Corporation New York.

Togninalli, M.; Ghisu, E.; Llinares-López, F.; Rieck, B.; and Borgwardt, K. 2019. Wasserstein Weisfeiler-Lehman Graph Kernels. *Advances in Neural Information Processing Systems*.