Efficient Rule Ensemble Learning using Hierarchical Kernels

Pratik Jawanpuria PRATIK.J@CSE.IITB.AC.IN
J. Saketha Nath SAKETH@CSE.IITB.AC.IN
Ganesh Ramakrishnan GANESH@CSE.IITB.AC.IN
Dept. of CSE, IIT-Bombay, Mumbai, INDIA.

Abstract

This paper addresses the problem of Rule Ensemble Learning (REL), where the goal is simultaneous discovery of a small set of simple rules and their optimal weights that lead to good generalization. Rules are assumed to be conjunctions of basic propositions concerning the values taken by the input features. From the perspectives of interpretability as well as generalization, it is highly desirable to construct rule ensembles with low training error, having rules that are i) simple, *i.e.*, involve few conjunctions and ii) few in number. We propose to explore the (exponentially) large feature space of all possible conjunctions optimally and efficiently by employing the recently introduced Hierarchical Kernel Learning (HKL) framework. The regularizer employed in the HKL formulation can be interpreted as a potential for discouraging selection of rules involving large number of conjunctions – justifying its suitability for constructing rule ensembles. Simulation results show that, in case of many benchmark datasets, the proposed approach improves over state-of-the-art REL algorithms in terms of generalization and indeed learns simple rules. Unfortunately, HKL selects a conjunction only if all its subsets are selected. We propose a novel convex formulation which alleviates this problem and generalizes the HKL framework. The main technical contribution of this paper is an efficient mirror-descent based active set algorithm for solving the new formulation. Empirical evaluations on REL problems illustrate the utility of generalized HKL.

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

1. Introduction

One of the most expressive and human readable representations for learned hypotheses is sets of if-then decision rules. A decision rule (Rivest, 1987) is a simple logical pattern of the form: *if condition then decision*. The condition consists of a conjunction of a small number of simple boolean statements (propositions) concerning the values of the individual input variables while the decision part specifies a value of the function being learned. The dominant paradigm for induction of rule sets, in the form of decision list (DL) models for classification (Rivest, 1987; Michalski, 1983; Clark & Niblett, 1989) has been a greedy *sequential covering* procedure.

Our work falls in the league of research (Weiss & Indurkhya, 2000; Cohen & Singer, 1999; Friedman & Popescu, 2008; Gao et al., 2007; Dembczyński et al., 2008; 2010) on Rule Ensemble Learning (REL). REL is a more general approach that treats decision rules as base classifiers in an ensemble. These models are additive in the rules that have optimized weights (coefficients). This is in contrast to the more restrictive DL models that are disjunctive sets of rules and which use only one in the set for each prediction. As pointed out in (Cohen & Singer, 1999), boosted rule ensembles are in fact simpler, better-understood formally than other state-of-the-art rule learners and also produce comparable predictive accuracy.

REL approaches like SLIPPER (Cohen & Singer, 1999), LRI (Weiss & Indurkhya, 2000), RuleFit (Friedman & Popescu, 2008), ENDER/MLRules (Dembczyński et al., 2010; 2008) have additionally addressed the problem of learning a compact set of rules in order to maintain its readability and also learn models that generalize better. Further, a number of rule learners like LRI, RuleFit encourage shorter rules (*i.e.*, fewer conjunctions in the condition part of the rule) or rules with a restricted number of conjunctions, for purposes of interpretation.We build upon this and define our Rule Ensemble Learning (REL) problem as that of

discovering a small set of simple rules and their weights that lead to good generalization. Some of these learners (Friedman & Popescu, 2008; Dembczyński et al., 2010) propose a regularized (empirical) risk minimization problem with a 1− norm regularization term on the weights of individual rules to encourage a compact set of non-trivial rules. However, all of them either approximate such a regularized solution using strategies such as shrinkage (in RuleFit, ENDER/MLRules) or resort to post-pruning (Gao et al., 2007). We provide efficient and optimal solutions to two regularized (empirical) risk minimization formulations for REL. The first is an application of Hierarchical Kernel Learning (Bach, 2009) (HKL), which discourages selection of rules involving large number of conjunctions and efficiently explores the exponentially large space of all possible conjunctions, in time polynomial in the number of selected rules, to yield a small set of simple rules. The second is a generalized formulation of HKL to further discourage glaring redundancies in the selected rule set. Again, for this generalized formulation, we develop an efficient algorithm that yields an optimal solution. Experimental results demonstrate the suitability of the proposed approaches for solving the REL problem. It is interesting to note that the generalized HKL approach, while yielding compact rule sets, achieved a 25% improvement in terms of generalization over state-of-the-art for a benchmark dataset.

We formally state the problem of Rule Ensemble Learning (REL) in Section 2. We note that the size of the space of conjunctions of simple boolean statements concerning values of the individual input variables is exponential in the number of the basic propositions. In Section 3 we solve the REL problem efficiently by posing it as an instance of HKL, which provides for efficient and optimal exploration of this exponentially large search space. Although the HKL formulation encourages selection of a small set of simple rules, it can be shown that if a rule with a particular condition is selected, then every subset of the conjunctions in the condition appears as the condition statement of some other selected rule. Such redundancies can hamper the readability of the rule-set and more importantly contradict our goal of selecting as few rules as possible. In Section 4, we propose a novel convex formulation that not only addresses this problem but also generalizes the HKL framework. We develop an efficient mirror-descent (Ben-Tal & Nemirovski, 2001) based active set algorithm for solving the new formulation. We report empirical evaluations of our HKL and generalized HKL formulations for REL on several publicly available datasets in Section 5. We also compare our results against several state-of-the-art decision list and rule ensemble learners. Results clearly illustrate that without compromising on accuracy, the HKL formulation yields shorter rules than state-of-the-art while generalized HKL additionally yields highly compact rule-sets. We summarize and conclude in Section 6.

2. Rule Ensemble Learning

Let $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ be the training data where $\mathbf{x}_i = [x_{i1}, \ldots, x_{in}]^\top \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$ represent the i^{th} input data point and the corresponding label respectively. Let $\mathcal{V}$ be a set of indices for the ensembles. Then the prediction model for ensemble learning takes the form (Friedman & Popescu, 2008): $\sum_{v\in\mathcal{V}} f_v \phi_v(\mathbf{x}) - b$ where, $\mathbf{x}$ is the input data point at which the prediction is made, $\phi_v(\cdot)$ is the v^{th} ensemble function and f_v is the weight given to this ensemble function. Ensemble learning algorithms differ based on the choice of the ensemble function class and on the learning algorithms employed for deriving the ensemble functions and their weights from the data.

Following the works on REL (Friedman & Popescu, 2008; Dembczyński et al., 2010), we assume that each ensemble $\phi_v(\mathbf{x})$, $v \in \mathcal{V}$ is a conjunction of basic propositions concerning the input feature values of $\mathbf{x}$; thus, $\phi_v(\mathbf{x}) : \mathbb{R}^n \to \{0, 1\}$. For a nominal input feature, say $j \in \{1, 2, \ldots, n\}$, and taking nominal values from the set $\{a_{j1}, \ldots, a_{jn_j}\}$, the basic propositions evaluated at a data point $\mathbf{x}$ take one of the forms: $x_j = a_{jk}$ or $x_j \neq a_{jk}$ for all $k = 1, \ldots, n_j$. For a numerical input feature, say $j \in \{1, 2, \ldots, n\}$, we pick n_j critical points, say $a_{j1}, \ldots, a_{jn_j}$. The basic propositions in this case are of one of the following forms: $x_j \leq a_{jk}$ and $x_j \geq a_{jk}$ for all $k = 1, \ldots, n_j$. To summarize, the total number of basic rules are $p = 2\sum_{j=1}^{n} n_j$. We consider $\mathcal{V}$ to be the set of all possible conjunctions of the p propositions. Then, $\phi_v(\mathbf{x})$ is the v^{th} conjunction of the basic rules evaluated on $\mathbf{x}$. Denote by $\mathbf{f}$ and $\phi(\cdot)$ the vectors with entries as f_v and $\phi_v(\cdot)$ for all $v \in \mathcal{V}$ respectively. The REL problem can be posed as that of determining the feature weights that minimize a weighted combination of the empirical risk[1] and a regularization term[2]: $\min_{\mathbf{f},b} \frac{1}{2}\left(\Omega(\mathbf{f})\right)^2 + C\sum_{i=1}^{m} l(y_i, \langle \mathbf{f}, \phi(\mathbf{x})\rangle - b)$ where C is the regularization parameter and l is a convex

[1] Here, $\langle \mathbf{u}_1, \mathbf{u}_2 \rangle$ denotes the inner product of $\mathbf{u}_1, \mathbf{u}_2$.

[2] Note that $\mathcal{V}$ as defined above may include some impossible conjunctions. This inclusion is done only for convenience sake. With any norm-based regularizer, it is easy to see that the weight f_v given to such conjunctions is always zero. As we shall understand later, the complexity of the proposed active set algorithm is not adversely affected by this.

loss function (e.g., hinge loss, logistic loss, huber loss etc.). The regularization term (Friedman & Popescu, 2008; Dembczyński et al., 2010), $(\Omega(\mathbf{f}))^2 = \|\mathbf{f}\|_1^2$, is meant to encourage a sparse $\mathbf{f}$, so that only a few rules are learned. However, to the best of our knowledge, rule ensemble learners that identify the need for sparse $\mathbf{f}$, either approximate the solution or perform post-pruning (Friedman & Popescu, 2008; Dembczyński et al., 2010; Gao et al., 2007). Moreover, as motivated earlier, it is desirable to learn rules which are simple *i.e.*, involving as few rules as possible and for this Ω needs to be chosen appropriately.

In this paper, we propose to solve this problem optimally for a family of functions $\Omega(\mathbf{f})$ which discourage selection of large conjunctive rules. Note that $\mathcal{V}$ has a size exponential in p, the number of basic rules. Thus, our algorithm should also efficiently search this exponentially large space. To achieve this end, we note that $\mathcal{V}$ has a structure; it can be easily verified that $\langle \mathcal{V}, \subseteq \rangle$ is a lattice and will be hereafter referred to as the *conjunction lattice*. Here $\forall\ v_1, v_2 \in \mathcal{V},\ v_1 \subseteq v_2$ *iff* v_1 is a subset of the conjunctions of v_2. In the next section, we leverage the recent work of Hierarchical Kernel Learning (Bach, 2009) to explore this structure in $\mathcal{V}$ and solve the REL problem in time polynomial in the number of selected rules.

Some more notation is in order before we move on. We follow the convention that the top (level 0) of the Hasse diagram depicting the *conjunction lattice* corresponds to the empty conjunction and the bottom (level p) to the conjunction of all basic propositions. Under this convention, the nodes at level 1, say, denoted by B, form the set of basic propositions. Let $D(v)$ and $A(v)$ represent the set of descendants and ancestors of the node v in the lattice. We assume that both $D(v)$ and $A(v)$ include the node v. For any subset of nodes $\mathcal{W} \subset \mathcal{V}$, the hull and sources of $\mathcal{W}$ are defined as: $hull(\mathcal{W}) = \bigcup_{w \in \mathcal{W}} A(w)$ and $sources(\mathcal{W}) = \{w \in \mathcal{W} \mid A(w) \cap \mathcal{W} = \{w\}\}$. Also, $|\mathcal{W}|$ denotes the size of the set and $\mathcal{W}^c$ the denotes the complement *i.e.*, all the nodes in $\mathcal{V}$ which are not in $\mathcal{W}$. We let $\mathbf{f}_{\mathcal{W}}$ denote the vector with entries $f_v,\ v \in \mathcal{W}$. In general, the entries in a vector are referred to using an appropriate subscript *i.e.*, entries in $\mathbf{u}_i \in \mathbb{R}^d$ are denoted by $u_{i1}, \ldots, u_{id}$ etc.

3. Rule Ensemble Learning using HKL

One way of discouraging selection of conjunctions involving many basic rules is by employing a $(1,2)$ mixed-norm based regularizer promoting group sparsity: $\Omega(\mathbf{f}) = \sum_{v \in \mathcal{V}} \|\mathbf{f}_{D(v)}\|_2$. Since the 1-norm is known to promote sparsity (Rakotomamonjy et al., 2008; Bach, 2009), for most of the $v \in \mathcal{V}$, we have that $\|\mathbf{f}_{D(v)}\|_2 = 0$ ($f_w = 0\ \forall\ w \in D(v)$). Hence the selection of conjunctions near the bottom of the *conjunction lattice*, which correspond to those involving many basic rules, is indeed discouraged. In order to facilitate incorporation of prior information regarding the usefulness of the feature conjunctions, one may also employ a weighted version of the above regularizer: $\Omega_H(\mathbf{f}) = \sum_{v \in \mathcal{V}} d_v \|\mathbf{f}_{D(v)}\|_2$ where $d_v \geq 0$ is a non-negative weight parameter. This leads to the following hierarchical kernel learning formulation (Bach, 2009):

$$\min_{\mathbf{f}, b} \frac{1}{2} \left(\Omega_H(\mathbf{f})\right)^2 + C \sum_{i=1}^{m} \left[l\left(y_i, \langle \mathbf{f}, \phi(\mathbf{x}_i) \rangle - b\right)\right] \qquad (1)$$

where $\Omega_H(\mathbf{f}) = \sum_{v \in \mathcal{V}} d_v \|\mathbf{f}_{D(v)}\|_2$.

HKL was introduced as a generic framework for performing non-linear feature selection. In the context of the HKL formulation (1), $\mathcal{V}$ can be a set of nodes in any directed acyclic graph and need not be restricted to those in the *conjunction lattice*; *i.e.*, $\phi_v(\cdot)$ is not restricted to be a boolean value, but may be any vector in the RKHS induced by a kernel, say k_v *i.e.*, $k_v(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi_v(\mathbf{x}_i), \phi_v(\mathbf{x}_j) \rangle$. In this case f_v must itself be a vector in this RKHS and $\langle \mathbf{f}, \phi(\mathbf{x}) \rangle = \sum_{v \in \mathcal{V}} \langle f_v, \phi_v(\mathbf{x}) \rangle$. Accordingly, $\mathbf{f}_{D(v)}$ is now defined as the vector with entries given by $\|f_w\|_2\ \forall\ w \in D(v)$. Note that this generalization in the notation is indeed consistent with that introduced previously in the context of REL — where $\mathcal{V}$ represents the set of nodes in the conjunction lattice and $\phi_v(\cdot)$ and f_v are one dimensional. In the remainder of this paper we follow this generalized notation while keeping in mind the specialization in the context of REL.

The dual of (1) turns out to be the problem of learning sparse (non-negative) linear combinations of the exponentially large number of kernels k_v. Such problems have also been studied in the name of Multiple Kernel Learning (MKL) (Rakotomamonjy et al., 2008). Typically in the setting of MKL problems, few base kernels are given and the goal is to optimally combine them in order to achieve good generalization. Thus HKL formulation is an extreme case of MKL where the number of base kernels is exponential. Since the RKHS induced by sum of kernels is the concatenation of those induced by the individual kernels, selection of kernels is equivalent to selection of the corresponding induced RKHS/feature spaces. More specifically, in the context of REL, selection of kernels is equivalent to selection of conjunctive rules. The mixed-norm regularizer $\Omega_H(\mathbf{f})$ employed in the HKL formulation

was introduced by the authors from a purely computational perspective – in order to restrict the sparsity patterns such that an active set algorithm (refer figure 6 in Bach (2009)) can be used to efficiently solve the formulation. In fact, a key result in Bach (2009) is that the computational complexity of the active set algorithm is polynomial in the number of selected kernels — provided the kernels $k_v(\cdot,\cdot)$ are such that they can be summed over the descendants of a given node in polynomial time in the number of input features.

Note that, in our case of REL, the kernels do satisfy this criterion: the RKHS in which $\phi_v(\cdot)$ lies is essentially the one dimensional Euclidean space induced by the v^{th} conjunction and $k_v(\mathbf{x}_i, \mathbf{x}_j)$ is simply the product of the v^{th} conjunction evaluated on the examples $\mathbf{x}_i$ and $\mathbf{x}_j$. Since the basic rules are boolean, $k_v(\mathbf{x}_i, \mathbf{x}_j)$ is further equal to the product of evaluations of all the basic rules corresponding to the v^{th} conjunction on these two examples. It is easy to see that in this case the sum of these products over any sublattice formed by the descendants can be computed efficiently by re-writing it as product of sums; for e.g., $\sum_{v\in\mathcal{V}} k_v(\mathbf{x}_i, \mathbf{x}_j) = \Pi_{k\in B}(1+\phi_k(x_i)\phi_k(x_j))$ (recall that B is the set of level 1 nodes in the lattice *i.e.*, the basic rules themselves). Hence the active set algorithm (figure 6 in Bach (2009)) can be employed for efficiently solving the HKL formulation in the context of REL — enabling us to explore the exponentially large space of conjunctive rules in polynomial time. Empirical results in Section 5 show that the rule ensembles constructed using the HKL framework indeed learn simple conjunctive rules with better generalization in many datasets.

Although these results are encouraging, it must be noted that under very mild conditions on the kernels, it can be shown that the HKL algorithm selects a conjunction only after selecting all conjunctions[3] which are subsets of it (refer theorem 6 in Bach (2009)). This, particularly in the context of REL, is psycho-visually redundant, because a rule with k propositional statements, if included in the result, will necessarily entail the inclusion of 2^k more general rules in the result. This violates the important requirement for a small set (Friedman & Popescu, 2008; Gao et al., 2007; Dembczyński et al., 2010) of human-readable rules (*c.f.* Section 1).

The key reason for this restrictive scheme of rule/feature selection in HKL is the presence of the 2-norm (which is known to promote non-sparse solutions) over the feature weights of the descendants (*i.e.*, $\|\mathbf{f}_{D(v)}\|_2$). To achieve our desired objective, we generalize the HKL formulation and employ the regularizer[4]: $\Omega_S(\mathbf{f}) = \sum_{v\in\mathcal{V}} d_v \|\mathbf{f}_{D(v)}\|_\rho,\ \rho \in (1, 2]$. Since the 1-norm promotes sparsity, as in case of the original HKL formulation, for most of $v \in \mathcal{V}$ we have that $\|\mathbf{f}_{D(v)}\|_\rho = 0$ ($f_w = 0\ \forall\ w \in D(v)$). But now, even in cases where $\|\mathbf{f}_{D(v)}\|_\rho$ is not forced to zero by the 1-norm, many of the feature weights of its descendants (*i.e.*, f_w for $w \in D(v)$) will be forced to zero as it is known that norms between $(1, 2)$ promote sparse solutions (for e.g., (Szafranski et al., 2008)). This translates into relaxing the restrictive scheme of feature selection in HKL. The details of the generalized HKL formulation and an efficient algorithm for solving it are presented in the subsequent section.

4. Generalized HKL Formulation

In this section we present the main technical contribution of the paper — the generalized HKL formulation and an efficient mirror-descent based active set algorithm for solving it. In order to keep the expressions simple, in the remainder of this paper we focus on the case of binary classification (labels $y_i \in \{-1, 1\}$) and present the details using the hinge-loss as the loss function l. It is easy to extend the following results for other learning problems with appropriate choices of loss functions and kernels.

As discussed in the previous section, we generalize the HKL formulation by replacing the 2-norm with a ρ-norm ($\rho \in (1, 2]$) over the weights of the descendants:

$$\begin{aligned} \min_{\mathbf{f},b,\xi} \quad & \tfrac{1}{2}\left(\textstyle\sum_{v\in\mathcal{V}} d_v \|\mathbf{f}_{D(v)}\|_\rho\right)^2 + C\sum_{i=1}^m \xi_i \\ \text{s.t.} \quad & y_i\left(\sum_{v\in\mathcal{V}} \langle f_v, \phi_v(\mathbf{x}_i)\rangle - b\right) \geq 1-\xi_i, \xi_i \geq 0 \end{aligned} \tag{2}$$

To the best of our knowledge such generalizations of HKL have not been studied in the literature. In the following text, we present an efficient mirror-descent based active set algorithm for solving this (exponentially large) convex problem (2). The computational complexity of this algorithm is polynomial in the size of the hull of the selected kernels.

We begin by making the following observation: the problem (2) remains the same when solved with the original set of variables (*i.e.*, $\mathbf{f}, b, \xi$) or when solved with only those $f_v \neq 0$ at optimality and b, ξ. However the computational effort required in the later case can be far lower, as it is expected that most of the variables f_v will be zero at optimality. This motivates us to

[3] This includes the empty conjunction as well.

[4] We, on purpose do not wish to include the case $\rho = 1$ as then there will be no hope of solving it in polynomial time.

Algorithm 1 Active Set Algorithm

Input: Training data $\mathcal{D}$, Oracle for computing k_v, Maximum tolerance (ϵ).
Initialize $\mathcal{W}$ as set containing only the top node.
Compute η, α by solving (7) using mirror-descent.
while suff. cond. (6) is not met **do**
  Add suff. cond. (6) violating nodes to $\mathcal{W}$
  Recompute η, α by solving (7)
end while
Output: $\mathcal{W}, \eta, \alpha$

explore an active set algorithm, which is similar in spirit to that in (Bach, 2009).

The active set algorithm (refer algorithm 1 for details) starts with an initial guess for the set of non-zero f_v. This set is referred to as the active set ($\mathcal{W}$). At each iteration, the following problem, which is the same as (2) with variables restricted to the active set, is solved:

$$\begin{aligned} \min_{\mathbf{f},b,\xi} \quad & \tfrac{1}{2}\left(\textstyle\sum_{v\in\mathcal{W}} d_v \|\mathbf{f}_{D(v)\cap\mathcal{W}}\|_\rho\right)^2 + C\sum_{i=1}^m \xi_i \\ \text{s.t.} \quad & y_i\left(\sum_{v\in\mathcal{W}} \langle f_v, \phi_v(\mathbf{x}_i)\rangle - b\right) \geq 1-\xi_i, \xi_i \geq 0 \end{aligned} \quad (3)$$

If the optimal solution for this small problem (3), henceforth referred to as the *reduced solution*, is an optimal solution for the original problem (2), then the algorithm terminates; else the active set is updated accordingly. The two issues which need to be resolved are: i) can the optimality of the reduced solution be verified efficiently? ii) can the small problem (3) be solved efficiently?

We begin by addressing the former issue regarding optimality of the reduced solution and present a sufficient condition for optimality which can be verified in polynomial time in the size of the active set. The sufficient condition follows from the insight given by a dual of (2) which is presented below. The problem (2) aims at finding the optimal vector in an RKHS. In order to facilitate the application of a suitable representer theorem for writing down the dual, we employ here a variational characterization of $\Omega_S(\mathbf{f})$. With repeated application of lemma 26 in Micchelli & Pontil (2005), it can be shown that[5]:

$$\Omega_S(\mathbf{f})^2 = \min_{\gamma\in\Delta_{|\mathcal{V}|,1}} \min_{\lambda_v\in\Delta_{|D(v)|,\hat{\rho}}\forall v\in V} \sum_{w\in\mathcal{V}} \delta_w^{-1}(\gamma,\lambda)\|f_w\|_2^2$$

where $\Delta_{d,r} = \left\{\eta\in\mathbb{R}^d \mid \eta\geq 0, \sum_{i=1}^d \eta_i^r = 1\right\}$, $\delta_w^{-1}(\gamma,\lambda) = \sum_{v\in A(w)} \frac{d_v^2}{\gamma_v\lambda_{vw}}$ and $\hat{\rho} = \frac{\rho}{2-\rho}$. Using this variational characterization and applying the representer theorem (see also Rakotomamonjy et al. (2008)), we derive a partial-dual (dual wrt. variables $\mathbf{f}, b, \xi$ alone) of (2):

$$\min_{\gamma\in\Delta_{|\mathcal{V}|,1}} \min_{\lambda_v\in\Delta_{|D(v)|,\hat{\rho}}\forall v\in\mathcal{V}} \max_{\alpha\in S(\mathbf{y},C)} G(\gamma,\lambda,\alpha) \quad (4)$$

where

$$G(\gamma,\lambda,\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2}\alpha^\top\left(\sum_{w\in\mathcal{V}} \delta_w(\gamma,\lambda)\mathbf{K}_w\right)\alpha$$

and $S(\mathbf{y},C) = \{\alpha\in\mathbb{R}^m \mid 0\leq\alpha\leq C,\ \sum_{i=1}^m y_i\alpha_i = 0\}$.

We now note the following key theorem which provides a sufficient condition for optimality of the reduced solution:

Theorem 1. *Suppose the active set $\mathcal{W}$ is such that $\mathcal{W} = hull(\mathcal{W})$. Let the reduced solution with this $\mathcal{W}$ be $(\mathbf{f}_\mathcal{W}, b_\mathcal{W}, \xi_\mathcal{W})$ and the corresponding dual variables be $(\gamma_\mathcal{W}, \lambda_\mathcal{W}, \alpha_\mathcal{W})$. The reduced solution is a solution of (2) with a duality gap less than ϵ if:*

$$\max_{t\in sources(\mathcal{W}^c)} \|\beta_t\|_{\bar{\rho}} \leq \left(\Omega_S(\mathbf{f}_\mathcal{W})\right)^2 + 2(\epsilon-\epsilon_\mathcal{W}) \quad (5)$$

where β_t is the vector with $|D(t)|$ number of entries and the w^{th} ($w\in D(t)$) entry being $\frac{\alpha_\mathcal{W}\mathbf{K}_w\alpha_\mathcal{W}}{\left(\sum_{v\in A(w)\cap D(t)} d_v\right)^2}$, $\bar{\rho} = \frac{\rho}{2(\rho-1)}$ and $\epsilon_\mathcal{W}$ is a duality gap term associated with the computation of the reduced solution.

The proofs of theorems 1 and 2 are fairly technical and not included here due to space restrictions. They are presented in Jawanpuria et al. (2011).

In the special case $\bar{\rho} = 1$ (*i.e.*, $\rho = 2$), this condition is same that obtained in Bach (2009) and can be verified efficiently: size of $sources(\mathcal{W}^c)$ is upper-bounded by $p|\mathcal{W}|$ and the sum of the quadratic terms $\alpha^\top\mathbf{K}_w\alpha$ can be done efficiently as we assumed that the kernels are easily summable over the descendants. Similarly, the terms involving the weights d_v also do not pose a problem as long as we choose the d_v such that they decompose as products. In this case $\sum_{v\in A(w)\cap D(t)} d_v$ can be computed in linear time in p. When $\rho\in(1,2)$ (*i.e.*, $\bar{\rho}\in(1,\infty)$), we still insist on $\max_{t\in sources(\mathcal{W}^c)}\|\beta_t\|_1 \leq$ RHS of (5). This will be fine because for any β_t, $\|\beta_t\|_{\bar{\rho}} \leq \|\beta_t\|_1$. At first, this may seem to be a pessimistic approach, however, in the context of REL problems, it is easy to see that each β_t will be a highly sparse vector as the most of the matrices $\mathbf{K}_w$, especially near the bottom of the lattice, will be (near) zero-matrices. This is because the larger the conjunctive rule, the fewer are the examples which may satisfy it. Thus effectively for all $\rho\in(1,2]$ we use the following sufficient condition:

$$\max_{t\in sources(\mathcal{W}^c)} \|\beta_t\|_1 \leq \left(\Omega_S(\mathbf{f}_\mathcal{W})\right)^2 + 2(\epsilon-\epsilon_\mathcal{W}) \quad (6)$$

[5] Refer Jawanpuria et al. (2011) for details.

In case the sufficiency criterion is satisfied at an iteration, the active set algorithm terminates; else the nodes in $sources(\mathcal{W}^c)$ which violate the condition are appended to the active set[6].

We now focus on the issue of solving the small problem (3) efficiently. Existing optimization tools like `cvx`[7] are capable of solving this problem; however our initial simulations showed that they are not scalable and impractical to work with real-world REL tasks. Also the wrapper approaches presented in (Szafranski et al., 2008; Bach, 2009) cannot be applied either to the primal form (3) or the dual form (4). The following important theorem presents a highly specialized dual of (2) which motivates a simple mirror-descent based algorithm:

Theorem 2. *The following is a dual of (2) and the objectives of (2), (4) and (7) are equal at optimality:*

$$\min_{\eta \in \Delta_{|\mathcal{V}|,1}} \quad g(\eta) \tag{7}$$

where $g(\eta)$ is the optimal objective value of the following convex problem:

$$\max_{\alpha \in S(\mathbf{y},C)} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\left(\sum_{w \in \mathcal{V}} \zeta_w(\eta)\left(\alpha^\top \mathbf{K}_w \alpha\right)^{\bar{\rho}}\right)^{\frac{1}{\bar{\rho}}} \tag{8}$$

where $\zeta_w(\eta) = \left(\sum_{v \in A(w)} d_v^\rho \eta_v^{1-\rho}\right)^{\frac{1}{1-\rho}}$ and $\mathbf{K}_w$ is the $m \times m$ matrix with entries as: $y_i y_j k_w(\mathbf{x}_i, \mathbf{x}_j)$.

Apart from motivating an efficient algorithm for solving the generalized HKL formulation, the dual (7) gives some valuable insights into the formulation; since (7) is essentially a 1-norm regularized problem, it is expected that most of the η's will be (near) zero at optimality. Also, the weight $\zeta_w(\eta)$ is zero whenever one or more of the $\eta_v (v \in A(w))$ are zero. In the context of REL, these observations imply that the formulation (2) discourages selection of kernels near the bottom of the lattice or equivalently large conjunctive rules.

Interestingly, the problem in (8) is equivalent to the $\hat{\rho}$-norm MKL (Kloft et al., 2009) problem with $\hat{\rho} = \frac{\rho}{2-\rho}$ and the base kernels are $(\zeta_v(\eta))^{\frac{1}{\bar{\rho}}} k_v \;\; \forall \; v \in \mathcal{V} \;\; \ni \zeta_v(\eta) \neq 0$. The $\hat{\rho}$-norm MKL formulation problem promotes sparsity in combining the base kernels whenever $\hat{\rho} < \infty$ (*i.e.*, $\rho \in (1,2)$) and highly sparse solutions when $\hat{\rho} < 2$ (*i.e.*, $\rho \in (1, 4/3)$) (Szafranski et al., 2008; Kloft et al., 2009). It can also be shown that if $\hat{\rho} = \infty$ (*i.e.*, $\rho = 2$), then all the base kernels will be selected[8]. Consider a node $w \in \mathcal{V}$ such that none of the η_v $(v \in A(w))$ are zero — which in turn implies that none of ζ_v, $v \in A(w)$ is zero. Then the above MKL interpretation implies that the kernel k_w will be selected if $\rho = 2$ and may not be selected whenever $\rho \in (1,2)$. In the context of REL this implies that conjunctions may be selected regardless of the selection of its subsets provided $\rho \in (1,2)$. This justifies our generalized HKL formulation.

The small problem (3) can be solved efficiently by solving the dual problem given by theorem 2 (*i.e.*, problem (7) with $\mathcal{V}$ restricted to $\mathcal{W}$). In the following we propose to employ the mirror-descent algorithm (Ben-Tal & Nemirovski, 2001) for solving (7). Mirror-descent is a variant of the projected (sub)gradient-descent algorithm and can be employed to solve any problem of the form: $\min_{\mathbf{x} \in X} h(\mathbf{x})$, where h is a convex, Lipschitz continuous function and X is a convex compact set. It is assumed that there exists an oracle which computes the (sub)gradient of h at any given $\mathbf{x} \in X$. Mirror-descent algorithms are known to efficiently solve such problems, especially those with feasibility set being a simplex. We note the following theorem which justifies the applicability of mirror-descent for solving (7):

Theorem 3. *The function $g(\eta)$ given by (8) is convex. Also, the i^{th} entry in the sub-gradient $(\nabla g(\eta))_i = -\frac{d_i^\rho \eta_i^{-\rho}}{2\bar{\rho}}\left(\sum_{w \in \mathcal{V}} \zeta_w(\eta)\left(\bar{\alpha}^\top \mathbf{K}_w \bar{\alpha}\right)^{\bar{\rho}}\right)^{\frac{1}{\bar{\rho}}-1}\left(\sum_{w \in D(i)} \zeta_w(\eta)^\rho \left(\bar{\alpha}^\top \mathbf{K}_w \bar{\alpha}\right)^{\bar{\rho}}\right)$ where $\bar{\alpha}$ is an optimal solution of problem (8) with that η where the sub-gradient is to be computed. If all the eigen-values of the gram-matrices $\mathbf{K}_w$ are finite and non-zero, then g is Lipschitz continuous.*

Proof. We begin by noting that $\zeta_v(\eta)$ is a concave function of η for all v (this is because when $\rho \in (1,2]$, ζ_v is a weighted q-norm in η, where $q \in [-1, 0)$ and hence is concave in the first quadrant). By simple observations regarding operations preserving convexity we have that the objective in (8) is a convex function of η for a fixed value of α. Hence $g(\eta)$, which is a point-wise maximum over convex functions, is itself convex. The sub-gradient follows from the Danskin's theorem (prop. B.25 in Bertsekas (1999)). The Lipschitz continuity of g follows from the boundedness of this sub-gradient[9]. □

It is clear from the above theorem that computing the gradient of h involves solving (8) which is, as noted

[6] It is easy to see that with this update scheme of the active set, $\mathcal{W}$ is always equal to $hull(\mathcal{W})$ and hence theorem 1 holds.

[7] Available at `cvxr.com/cvx`

[8] This requires that all gram-matrices with each k_v are positive definite and is hence consistent with the theorem 6 in Bach (2009).

[9] The details are fairly technical and presented in Jawanpuria et al. (2011).

before, equivalent to solving a $\hat{\rho}$-norm MKL problem. This can be done using cutting planes algorithms (Kloft et al., 2009). In the special case $\rho = 2$, (8) is simply a regular SVM problem.

In the following text, the computational complexity of the active set algorithm is estimated assuming R is the final active set size. We need to solve (7) $O(R)$ times. Each mirror descent run takes $\log(R)$ iterations (Ben-Tal & Nemirovski, 2001) with dominant computation at each iteration being that of solving $\hat{\rho}$-norm MKL whose conservative complexity estimate is $O(m^3R^2)$. This amounts to $O(m^3R^3\log(R))$. The cost for computing gram-matrices is $O(m^2Rp)$; whereas that of verifying the sufficient conditions is $O(m^2R^2p)$. Thus the overall computational complexity is: $O(m^3R^3\log(R) + m^2Rp + m^2R^2p)$.

5. Experimental Results

In this section, we report the results of our simulations on classification datasets from the UCI repository.The goal is to compare various rule ensemble learners on the basis of (a) generalization — measured by the predictive performance on unseen test data (b) interpretability — measured using i) number of conjunctive rules employed ii) average number of propositions per rule.

The following methods were compared. More details of the experimental setup can be found in Jawanpuria et al. (2011).

$\mathbf{HKL}_\rho$: The proposed REL algorithms described in Sections 3 & 4. We consider three different values of ρ : 2, 1.5 and 1.1. Note that with $\rho = 2$, the formulation is the same as (Bach, 2009).

$\mathbf{ENDER}_M$: State-of-the-art rule ensemble algorithm (Dembczyński et al., 2010). The number of rules (denoted by M) was set to 500. Additionally, to enable comparisons between ENDER and the proposed methods, we also considered $M = \theta = \max(N_{1.5}, N_{1.1})$, where N_ρ is the number of rules computed by $\mathbf{HKL}_\rho$).

SLI: The SLIPPER algorithm (Cohen & Singer, 1999). The parameter settings were as in Dembczyński et al. (2010).

RuleFit: Rule ensemble algorithm as in Friedman & Popescu (2008).The parameter settings were the same as those specified by the authors.

We experimented with several benchmark datasets from the UCI repository[10]. For each dataset, we created 5 random train-test splits with 10% train data[11].

[10]For multiclass datasets, we picked the two most populated classes for binary classification.

[11]Except for monk-3 where train-test split was given.

Since most datasets were highly unbalanced, we report the average F1-score[12] (with standard deviation) for each dataset in Table 4. We also report the number of rules produced and the average length (number of propositions) of the rules[13]. It is desirable that REL algorithms achieve high F1-score with a small set of simple rules *i.e.*, compact set of rules.

We observe in Table 4 that the generalization of the proposed $\mathbf{HKL}_\rho$ is atleast as good as that of state-of-the-art, with the additional advantage that $\mathbf{HKL}_\rho$ learns rules with small number of conjunctions. More interestingly, $\mathbf{HKL}_{1.1}$ yields extremely small to medium-sized accurate rule sets. Wilcoxon sign rank test shows that there is only a small evidence (probabilities of 0.027, 0.049 and 0.02 respectively) in favour of the null hypotheses that the median F1-score with $\mathbf{HKL}_{1.1}$ is the same as those with **RuleFit**, **SLI** and $\mathbf{ENDER}_\theta$. Some observations from the table are noteworthy. For instance, $\mathbf{HKL}_{1.1}$ benefits from its optimal search and chooses a moderate sized ruleset for the tic-tac-toe dataset leading to substantial gain in F1-score. Whereas, in cases like monk-3, vote and blood trans. $\mathbf{HKL}_{1.1}$ selects extremely compact rulesets while performing comparably or marginally better in terms of F1-score. As expected, the compactness in the ruleset rapidly increases as ρ decreases while sometimes even improving on the F1-score. In fact we observed that the percentage of non-selected rules in the final hull for different HKL methods are: $\mathbf{HKL}_2$ 0%, $\mathbf{HKL}_{1.5}$ 48.1% and $\mathbf{HKL}_{1.1}$ 86.8% — which substantiates our intuition behind generalizing HKL. It is also interesting to note that $\mathbf{HKL}_{1.1}$ generalizes better even with the default configuration in ENDER which allows it to learn up to 500 rules.

6. Conclusions

We posed the problem of learning compact rule ensembles as a hierarchical kernel learning problem. In order to overcome the redundancies in the rule-set induced by HKL, we generalized the 1-norm in HKL to ρ-norm. We developed an efficient algorithm to optimally solve this generalization. This generalization of HKL could prove useful in other machine learning applications as well. Experimental results clearly illustrate that the generalized HKL can indeed learn a small set of simple and accurate rules.

[12]In order to be comparable with other REL works we also report accuracies. The details appear in Jawanpuria et al. (2011)

[13]Specified below each F1-score as: (number of rules, average length of rules)

Table 1. F-score, number of rules and average length of rules

Dataset	RuleFit	SLI	ENDER_M		HKL_ρ		
			$M = 500$	$M = \theta$	$\rho = 2$	$\rho = 1.5$	$\rho = 1.1$
TIC-TAC-TOE	0.652 ± 0.068 (**40**, 2.51)	0.747 ± 0.026 (59, 2.35)	0.678 ± 0.014 (500, 2.67)	0.633 ± 0.011 (111, 2.46)	0.889 ± 0.029 (129, 1.85)	0.904 ± 0.039 (111, 1.83)	$\mathbf{0.935} \pm 0.043$ (79, **1.77**)
BALANCE	0.835 ± 0.034 (**17**, 2.18)	0.856 ± 0.027 (25, 1.88)	0.893 ± 0.017 (500, 2.00)	0.827 ± 0.013 (64, 1.99)	0.893 ± 0.027 (65, 1.65)	$\mathbf{0.899} \pm 0.022$ (64, 1.62)	$\mathbf{0.899} \pm 0.023$ (28, **1.23**)
HABERMAN	0.512 ± 0.072 (**6**, 1.68)	0.565 ± 0.066 (8, **1.14**)	0.585 ± 0.016 (500, 1.84)	0.424 ± 0.000 (18, 1.87)	$\mathbf{0.594} \pm 0.056$ (32, 1.27)	$\mathbf{0.594} \pm 0.056$ (18, 1.24)	$\mathbf{0.594} \pm 0.056$ (12, 1.20)
CAR	0.913 ± 0.033 (**34**, 3.12)	0.895 ± 0.024 (141, 2.27)	0.908 ± 0.019 (500, 2.67)	0.755 ± 0.028 (80, 1.85)	$\mathbf{0.943} \pm 0.024$ (87, 1.78)	0.937 ± 0.033 (80, 1.77)	0.935 ± 0.036 (50, **1.68**)
BLOOD TRANS.	0.549 ± 0.092 (18, 1.99)	0.559 ± 0.100 (**6**, **1.07**)	$\mathbf{0.616} \pm 0.059$ (500, 1.73)	0.489 ± 0.054 (58, 1.5)	0.594 ± 0.009 (242, 1.64)	0.593 ± 0.011 (58, 1.76)	0.593 ± 0.011 (7, 1.40)
CMC	0.632 ± 0.013 (39, 2.41)	0.601 ± 0.041 (**13**, 2.13)	0.651 ± 0.014 (500, 2.90)	0.644 ± 0.026 (74, 2.65)	0.656 ± 0.014 (217, 1.96)	0.652 ± 0.023 (74, 1.77)	$\mathbf{0.659} \pm 0.008$ (43, **1.70**)
MONK-3	0.330 (8, **1**)	0.910 (20, 2.12)	**0.972** (500, 2.57)	**0.972** (7, 1.57)	**0.972** (18, **1**)	**0.972** (7, **1**)	**0.972** (**2**, **1**)
VOTE	0.940 ± 0.060 (12, 1.5)	$\mathbf{0.970} \pm 0.003$ (**3**, **1**)	0.953 ± 0.030 (500, 1.12)	$\mathbf{0.970} \pm 0.003$ (8, **1**)	0.969 ± 0.003 (39, 1.25)	0.969 ± 0.003 (8, 1.16)	0.969 ± 0.003 (**3**, **1**)
BREAST-C	$\mathbf{0.550} \pm 0.049$ (10, 1.82)	$\mathbf{0.550} \pm 0.088$ (**7**, 1.33)	0.548 ± 0.040 (500, 1.87)	0.414 ± 0.000 (18, 1.83)	0.529 ± 0.048 (27, 1.15)	0.529 ± 0.048 (18, 1.16)	0.515 ± 0.049 (14, **1**)
MAM. MASS	$\mathbf{0.834} \pm 0.010$ (8, 1.98)	0.827 ± 0.010 (**4**, **1**)	0.832 ± 0.005 (500, 2.15)	0.830 ± 0.015 (12, 1.9)	0.815 ± 0.007 (17, 1.49)	0.815 ± 0.007 (12, 1.53)	0.815 ± 0.007 (5, 1.18)

Acknowledgments

We acknowledge Chiranjib Bhattacharyya for initiating discussions on optimal learning of rule ensembles.

References

Bach, F. High-Dimensional Non-Linear Variable Selection through Hierarchical Kernel Learning. Technical report, INRIA, France, 2009.

Ben-Tal, A. and Nemirovski, A. Lectures on Modern Convex Optimization: Analysis, Algorithms and Engineering Applications. *MPS/ SIAM Series on Optimization*, 1, 2001.

Bertsekas, D. *Non-linear Programming*. Athena Scientific, 1999.

Clark, Peter and Niblett, Tim. The cn2 induction algorithm. *Mach. Learn.*, 3(4):261–283, 1989.

Cohen, William W. and Singer, Yoram. A simple, fast, and effective rule learner. In *AAAI*, 1999.

Dembczyński, Krzysztof, Kotłowski, Wojciech, and Słowiński, Roman. Maximum likelihood rule ensembles. In *ICML*, 2008.

Dembczyński, Krzysztof, Kotłowski, Wojciech, and Słowiński, Roman. ENDER - a statistical framework for boosting decision rules. *DMKD*, 21(1):52–90, 2010.

Friedman, Jerome H. and Popescu, Bogdan E. Predictive learning via rule ensembles. *Annals of Applied Statistics*, 2:916–954, 2008.

Gao, Yang, Huang, Joshua Zhexue, and Wu, Lei. Learning classifier system ensemble and compact rule set. *Connect. Sci*, 19(4):321–337, 2007.

Jawanpuria, P., Nath, J. S., and Ramakrishnan, G. Efficient Rule Ensemble Learning using Hierarchical Kernels. Technical Report TR-CSE-2011-35, CSE, IIT-Bombay, 2011.

Kloft, M., Brefeld, U., Sonnenburg, S., Laskov, P., Muller, K.-R., and Zien, A. Efficient and Accurate p-Norm Multiple Kernel Learning. In *NIPS*, 2009.

Micchelli, Charles A. and Pontil, Massimiliano. Learning the Kernel Function via Regularization. *JMLR*, 6:1099–1125, 2005.

Michalski, Ryszard S. A theory and methodology of inductive learning. *Artif. Intell.*, 20(2):111–161, 1983.

Rakotomamonjy, A., Bach, F., Canu, S., and Grandvalet, Y. SimpleMKL. *JMLR*, 9:2491–2521, 2008.

Rivest, Ronald L. Learning decision lists. *Mach. Learn.*, 2(3):229–246, 1987.

Szafranski, M., Grandvalet, Y., and Rakotomamonjy, A. Composite Kernel Learning. In *ICML*, 2008.

Weiss, Sholom M. and Indurkhya, Nitin. Lightweight rule induction. In *ICML*, 2000.