3Q1-7in1

サンプリングを用いた機械学習パイプライン探索手法

Automated Machine Learning with Dataset Sampling

塩田 哲哉 *1　及川 一樹 *1　澤田 雅人 *1
Tetsuya Shioda　Kazuki Oikawa　Masato Sawada

*1日本電信電話株式会社 ソフトウェアイノベーションセンタ
NTT Software Innovation Center

Machine learning business applications are on the rise, however, the gap between data scientists supply and demands is getting bigger. To tackle the serious problem, many researchers and organizations develop automated machine learning tools that non-experts can use easily. Automated machine learning mainly aims to optimize pipelines, e.g. preprocessing selection, algorithm selection and hyperparameter optimization. Most of academic tools are based on bayesian optimization techniques but they sometimes fails to obtain fine prediction models because of the high-dimensional search space. In this paper, we propose a pipeline optimization methods that searches promising pipelines using sampling data to get predictive models in a short time. Experimental results show that our method obtains better predictive models than auto-sklearn does for 67% datasets and the same predictive models for remaining datasets.

1. はじめに

学術分野，産業分野問わずに様々な領域に機械学習を用いたデータ分析を適用しようとする事例は増加している [1]．しかし，データ分析には統計や機械学習の知識が必要となるため，非データ分析専門家が実用レベルの予測モデルを構築するためには中長期的な教育や経験が必要となるのが現状である．

非データ分析専門家でも機械学習を用いたデータ分析に簡単に取り組むことができるように，機械学習の自動化 (AutoML) に関する研究 [2, 3] が近年盛んに行われている．機械学習系の国際会議 ICML では 2014 年から毎年 AutoML Workshop が開催されており，並行して AutoML Challenge[4] というデータ分析大会が 2016 年に開催されている．

AutoML Challenge で継続的に上位入賞をしている state-of-the-art な手法の 1 つが auto-sklearn[3] である．auto-sklearn は SMAC[5] と呼ばれる Random Forest を用いた逐次最適化手法を用いて，前処理および予測アルゴリズムの組み合わせ，すなわちパイプラインを探索することで高精度な予測を実現する．しかし，文献 [6] によると，パイプライン探索タスクにおいて SMAC は Random Search と大幅な精度差が見られないということが報告されている．

一方，データ分析専門家が予測モデルを構築する工程に着目すると，前処理や予測アルゴリズムの組み合わせやハイパーパラメータなどを適宜変更しながら，交差検証などを用いて有効性を逐次確認し，精度が向上したならその設定を採用する，といった手順で分析を進めることがよく見られる．

本論文では，データ分析専門家の予測モデル構築工程に着目し，データセットをサンプリングして，前処理や予測アルゴリズムの有効性を効率的に逐次確認しながら，高精度な予測モデルを構築するためのパイプライン探索手法を提案する．

本論文の構成は以下のとおりである．2 章では一般的な予測モデルの構築方法について述べる．3 章では提案手法について述べる．4 章では提案手法の評価実験を行い，5 章で本論文のまとめを行う．

連絡先: 塩田哲哉，日本電信電話株式会社，〒 180-8585 東京都武蔵野市緑町 3-9-11，shioda.tetsuya@lab.ntt.co.jp

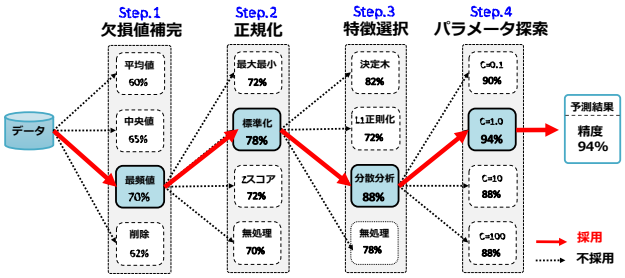


図 1: 予測モデルの構築方法の一例．各 Step 内の四角は前処理やパラメータ探索を実行したときの交差検証精度を示している．

2. 予測モデル構築の流れ

データ分析専門家が機械学習を用いて分類や回帰などの予測モデルを構築するための手順の一例を図 1 に示す．

始めにデータ分析専門家は達成したい目的にあわせてデータを収集する．例えば，優良顧客を発見するために過去の購買履歴やユーザ属性を収集する．

続いて，精度の高い予測モデルを構築できるように，収集したデータに対して様々な前処理を施す．前処理には欠損値補完や正規化といった様々な処理が存在し，欠損値補完という 1 つの処理に関しても，中央値補完や最頻値補完といった複数の手法が存在する．データ分析専門家は，交差検証法により予測精度が向上した手法を選択することにより，データに施すべき前処理を取捨選択し，機械学習パイプラインを組み立てていく．前処理の中には特徴量を何 % 削減するか，といったハイパーパラメータを含むものも存在するため，それらのハイパーパラメータもあわせて調整する必要がある．

最後に，前処理済みデータに対して，決定木や線形分類器といった複数の機械学習アルゴリズムを用いて予測モデルを構築する．予測精度が不十分であれば，機械学習アルゴリズムのハイパーパラメータの調整を行うことにより，予測精度を向上させる．

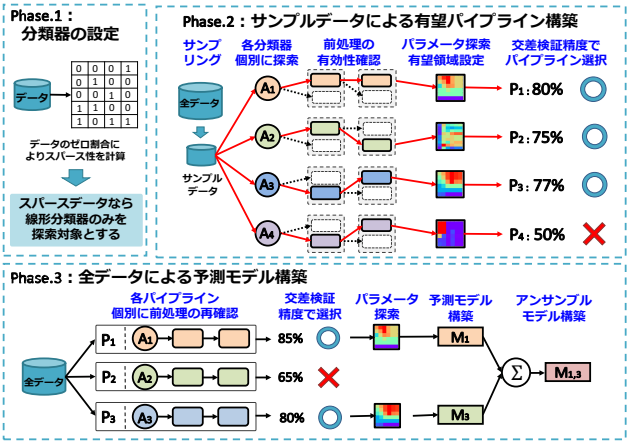


図 2: 提案手法の概要

上記の予測モデル構築方法を機械的に実装してパイプライン探索を行う場合，探索対象となる 前処理や機械学習アルゴリズムの数，および取りうるハイパーパラメータの数に比例して交差検証回数が増加する．一度の交差検証に要する時間は決して短いものではないため，愚直に全ての組み合わせを実行することは実行時間の観点で現実的ではない．一方で，前処理や機械学習アルゴリズムの有効性は実際に交差検証を行うまでわからないため，あらかじめ前処理や機械学習アルゴリズムを絞り込むことは予測精度の観点で望ましくない．

3. 提案手法

本章では，データ分析専門家の予測モデル構築を効率的に実現し，予測精度の高いモデルを構築する手法として，サンプリングを用いた機械学習パイプライン探索手法の提案を行う．3.1 節にて全体の枠組みについて説明し，3.2 節以降にて提案手法の詳細について述べる．なお，本論文では機械学習アルゴリズムの中で分類アルゴリズムを対象として手法を構築する．

3.1 概要

2. で述べた通り，前処理や分類器の有効性は交差検証を行うまでわからない一方で，愚直に全ての組み合わせを試行するのは実行時間の観点で現実的ではない．そこで，交差検証に要する時間はデータ数と正の相関があることを利用し，収集したデータをサンプリングして交差検証時間を短縮することにより，前処理や分類器の有効性を短時間で確認する．有効性が確認された前処理や分類器のみを対象にして全てのデータを用いて予測モデルを構築することにより，実行時間と予測精度を両立した探索を実現する．

提案手法の概要を図 2 に示す．本手法は分類器の設定，パイプライン構築，予測モデル構築の 3 フェーズに分けて分析を行う．Phase.1 の分類器の設定では，データの特性に応じて探索するべき分類器の設定を行う．Phase.2 のパイプライン構築では，与えられたデータをサンプリングし，各前処理，分類器の有効性を短時間で検証し，高精度な予測ができると期待されるパイプラインを複数構築する．Phase.3 の予測モデル構築では，Phase.2 で構築した各パイプラインに対して，全データを用いて前処理や分類器の有効性を再確認し，前処理に対する過適合を防いだ上で，予測モデルを構築する．最終的に構築された各予測モデルをアンサンブルし，さらなる予測精度の向上を狙う．以下では，本手法の詳細について述べる．

3.2 準備：予測精度の検証方法

本手法では交差検証とホールドアウト検証を併せて利用することにより，汎化性能の高い予測モデルを構築する．データセット $\mathcal{D}$ を交差検証用データとホールドアウト検証用データに分割したものをそれぞれ $\mathcal{D}_{\mathrm{cv}}$，$\mathcal{D}_{\mathrm{ho}}$ とする．Phase.2 のパイプライン構築では $\mathcal{D}_{\mathrm{cv}}$ からサンプリングしたデータを用い，Phase.3 の予測モデル構築では $\mathcal{D}_{\mathrm{cv}}$ を全て利用する．$\mathcal{D}_{\mathrm{ho}}$ は予測モデルの汎化性能を測定し，ユーザにベストモデルを提示するために利用する．

3.3 Phase.1: 分類器の設定

分類器は構築される識別境界の形状から線形分類器と非線形分類器に大別することができる．テキストのように非常にスパースなデータを分類する場合には，線形分類器は非線形分類器と同程度の精度で，なおかつ高速に予測を行うことができることが知られている [7]．そこで，データがスパースならば非線形分類器を探索対象から除外することで不要な探索を避け，実行時間を削減する．

スパース性判定にはデータに含まれるゼロの割合を用いる．各特徴量についてゼロの割合が閾値 r_f 以上ならば，その特徴量はスパースであると判断する．全特徴量に対するスパース特徴量の割合が閾値 r_a 以上ならば，そのデータはスパースであると判断する．

例として，探索対象の分類器がロジスティック回帰，決定木，k 近傍法の場合，データがスパースならばロジスティック回帰のみを探索対象とすることにより，実行時間を削減する．

3.4 Phase.2: パイプライン構築

ユーザパラメータであるサンプル率 s により $\mathcal{D}_{\mathrm{cv}}$ からサンプルデータ $\mathcal{D}^{(s)}_{\mathrm{cv,sample}}$ を抽出することにより，一度の交差検証時間を短縮する．探索対象の各分類器毎に，サンプルデータを用いて各前処理およびハイパーパラメータの有効性を確認し，高精度な予測が期待されるパイプラインを短時間で複数構築する．各分類器を A_j $(j = 1, 2, ...)$ と表記する．

3.4.1 前処理の探索

図 1 に示すように，各ステップごとに前処理の有効性を交差検証により確認し，最も予測精度が向上した前処理を採用して次のステップへ移行する．予測精度が向上しなかった場合には，その前処理を採用しない．ここで，前処理がハイパーパラメータを含む場合，グリッドサーチにより全てのハイパーパラメータの組み合わせを網羅的に探索することは多くの計算時間を要するため、各軸探索により最も予測精度の高いハイパーパラメータを探索する．

3.4.2 分類器のハイパーパラメータ探索・有望領域の設定

前処理済みデータに対して，各分類器 A_i のハイパーパラメータをグリッドサーチにより探索する．探索対象となったハイパーパラメータ集合を $\{\boldsymbol{\theta}_i^{A_j} \mid i = 1, 2, ...\}$ とし，各ハイパーパラメータに対する交差検証精度を $cv(\boldsymbol{\theta}_i^{A_j})$ で表し，その中で最大の交差検証精度を持つハイパーパラメータを $\boldsymbol{\theta}_{\max}^{A_j}$ で表す．

Phase.3 の予測モデル構築において，全データでハイパーパラメータ探索を行うことは計算コストが大きい．そのため，グリッドサーチの結果により高精度な予測が期待されるハイパーパラメータ探索空間の有望領域を設定する．各ハイパーパラメータの交差検証精度 $cv(\boldsymbol{\theta}_i^{A_j})$ について，「$cv(\boldsymbol{\theta}_{\max}^{A_j})$ と $cv(\boldsymbol{\theta}_i^{A_j})$ は等しい」という帰無仮説を立て，Welch の t 検定を行う．帰無仮説が棄却されない場合，$\boldsymbol{\theta}_i^{A_j}$ を有望なハイパーパラメータ $\boldsymbol{\theta}_{i,\mathrm{pm}}^{A_j}$ とする．有望なハイパーパラメータ集合 $\{\boldsymbol{\theta}_{i,\mathrm{pm}}^{A_j}\}$ の各次

元の最小値および最大値から構成される領域を探索空間の有望領域として，Phase.3 での全データを用いたハイパーパラメータ探索に利用する．

3.4.3 有望パイプラインの構築および選択

各分類器 A_j に対して，3.4.1 節で採用された前処理によりデータ変換を行い，3.4.2 節で採用されたハイパーパラメータ $\boldsymbol{\theta}_{\max}^{A_j}$ を用いて予測モデルを構築するパイプラインを $\boldsymbol{p}_j$ と表す．

Phase.3 の全データでの予測モデル構築は，サンプルデータよりも多くの時間が必要であるため，予測精度の高い有望なパイプラインのみを選択して Phase.3 での予測モデル構築対象とする．各パイプライン $\boldsymbol{p}_j$ の交差検証精度 $cv(\boldsymbol{p}_j)$ に対して，最大の交差検証精度との比率がある閾値 L 以上となった場合，すなわち

$$\frac{cv(\boldsymbol{p}_j)}{\max_j cv(\boldsymbol{p}_j)} \geq L \ , \ (0 \leq L \leq 1) \tag{1}$$

を満たした場合，$\boldsymbol{p}_j$ を有望パイプライン $\boldsymbol{p}_j^{\mathrm{pm}}$ とする．

3.5 Phase.3: 予測モデル構築

Phase.2 で得られた有望なパイプライン集合 $\{\boldsymbol{p}_j^{\mathrm{pm}}\}$ に対して，全データを用いて各前処理およびハイパーパラメータの有効性を再確認することで，前処理に対する過適合を防いだ上で，予測モデルを構築する．

3.5.1 前処理の有効性の再確認

サンプルデータに対する前処理の過適合を避けるために，全データ $\mathcal{D}_{\mathrm{cv}}$ を用いて各パイプライン $\boldsymbol{p}_j^{\mathrm{pm}}$ の前処理の有効性を交差検証により再確認し，予測精度が向上した前処理のみを採用する．

3.5.2 有望パイプラインの選択

分類器のハイパーパラメータ探索は交差検証が複数回実行されるため多くの時間を要する処理である．そのため，3.4.2 節で行ったパイプライン選択に加えて，更にハイパーパラメータ探索の対象となる有望なパイプラインを予め絞り込むことにより，無駄な交差検証を削減する．前処理の有効性確認済みの各パイプライン $\boldsymbol{p}_j^{\mathrm{pm}}$ について，3.4.3 節と同様に，式 (1) 交差検証精度の比率により，有望なパイプラインを選択する．選択された有望なパイプラインの集合を $\{\boldsymbol{p}_j^{\mathrm{pm}'}\}$ と表す．

3.5.3 分類器のハイパーパラメータ探索

選択された有望なパイプライン $\{\boldsymbol{p}_j^{\mathrm{pm}'}\}$ に対してハイパーパラメータ探索を行う．探索は 3.4.2 節で設定したハイパーパラメータの有望領域内で行われる．探索手法は任意であるが，本論文では TPE: Tree Parzen Estimators[8] を用いる．TPE は逐次的最適化法の一種であり，探索と活用をバランスよく行う最適化手法である．TPE は利用者が停止するまで探索を続けるため，交差検証精度の最大値が I_{stag} 回数更新されなかった場合，もしくは探索回数が $I_{\max}$ に達した場合に探索を打ち切る．

交差検証精度が最大となったハイパーパラメータ探索を用いて各パイプライン $\boldsymbol{p}_j^{\mathrm{pm}'}$ から予測モデル M_j を構築し，ホールドアウト検証用データ $\mathcal{D}_{\mathrm{ho}}$ を用いてホールドアウト検証精度 $ho(M_j)$ を測定する．

3.5.4 アンサンブル予測モデル構築

予測モデルの集合 $\{M_j\}$ をホールドアウト検証精度のよい順から，各ラベルに対する予測確率 $\boldsymbol{f}(\boldsymbol{x} \mid M_j)$ に重み付けを行うことにより，アンサンブル予測モデルを構築する．

$$\boldsymbol{f}(\boldsymbol{x} \mid e) = \sum_{j=1}^{e} w_i \boldsymbol{f}(\boldsymbol{x} \mid M_j) \tag{2}$$

表 1: 提案手法の前処理およびハイパーパラメータ一覧.

Step	前処理	#params	#cat	#cont
1	単一特徴量削除	0	0	0
2	欠損値補完	3	3	0
3	カテゴリ変数変換	0	0	0
4	スケール変換	4	4	0
5	特徴選択	7	7	0
6	次元削減	1	1	0
7	カーネル写像	2	2	0

表 2: 提案手法の分類器およびハイパーパラメータ一覧

分類器	線形	#params	#cat	#cont
Logistic Regression	O	3	2	1
Linear SVM	O	2	1	1
Random Forest	X	3	1	2
Extra Trees	X	3	1	2
k-Nearest Neighbors	X	2	1	1
SVM	X	4	2	2
Naive Bayes	X	0	0	0

ここで重み $\boldsymbol{w}$ は予測精度を最大化するようにランダムサーチにより設定する．

3.5.5 ベストモデルの選択

これまで構築した単体の予測モデルおよびアンサンブル予測モデルの中で，ホールドアウト検証精度が最良のモデルを選択し，ユーザに提示する．

4. 実験と評価

本節では，予測精度の観点で state-of-the-art な機械学習の自動化ツール auto-sklearn と提案手法の比較実験を行う．

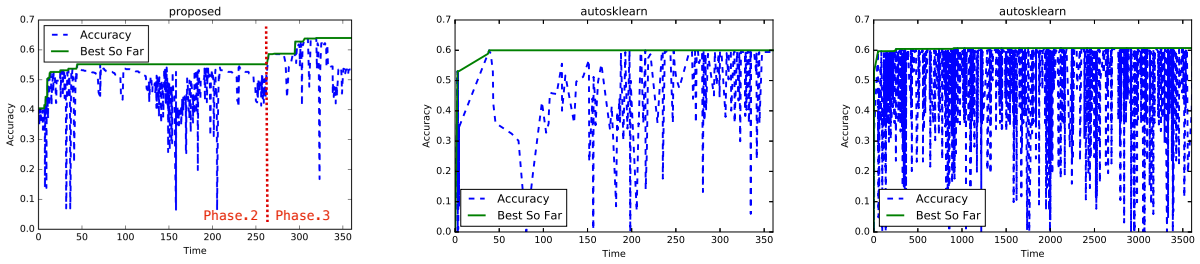
4.1 実験設定

auto-sklearn はユーザが打ち切るまで探索を続け，その中で最も交差検証精度の優れた予測モデルをユーザに提示する．そこで，提案手法と auto-sklearn を同一時間だけ実行したときの予測精度を用いて評価を行う．評価用データセットとして，UCI Machine Learning Repository[9] に公開されているデータを用いる．実験は 10 試行行い，テストデータに対する正解率の中央値により比較を行う．提案手法のユーザパラメータについて、サンプル率を $s = 0.2$，スパース判定閾値を $r_f = r_a = 0.9$，パイプライン選択閾値を $L = 0.9$，TPE の探索条件を $I_{\mathrm{stag}} = 5$，$I_{\max} = 20$，アンサンブル数を $e \in [2, 7]$ とする．利用する前処理および分類器一覧を表 1, 2 に示す．auto-sklearn は GitHub より取得した 0.1.1 版を用いる [*1]．

4.2 実験結果

実験結果を表 3 に示す．提案手法は auto-sklearn に対して 10/12 のデータセットで高い予測精度をもつ予測モデルを構築できていることがわかる．提案手法に対して auto-sklearn の予測精度が悪い傾向にある理由の 1 つとしては，auto-sklearn の探索空間の広さがあると考えられる．auto-sklearn は前処理および予測モデルのハイパーパラメータを一括して SMBO に基づき最適化をしているが，その探索空間は 100 次元を超える．SMBO は探索の初期段階において獲得関数に従いハイパーパラメータ空間を網羅的に探索し，その後，精度が高いと

∗1 https://github.com/automl/auto-sklearn/tree/0.1.1



(a) 提案手法 (360 秒)　(b) auto-sklearn (360 秒)　(c) auto-sklearn (3600 秒)

図 3: winequality データセットにおける交差検証精度の推移．提案手法は Phase.2 ではサンプルデータの， Phase.3 では全データの交差検証精度を示している．

表 3: 提案手法と auto-sklearn の正解率の比較．太字は最良の結果を示し下線は統計的有意差がないことを示す．

データセット	提案手法	auto-sklearn
car	**100**	99.04
dexter	**94.44**	92.78
dorothea	94.78	94.64
germancredit	73.17	73.48
krvskp	**99.58**	98.96
madelon	**88.14**	81.79
mnist	**96.60**	95.53
secom	**92.13**	91.70
semeion	93.51	93.61
waveform	**85.80**	85.43
winequality	**66.85**	63.65
yeast	61.12	61.01

見込まれる領域を重点的に探索する．そのため，文献 [6] においても述べられているように，限られた時間内で高次元のハイパーパラメータ空間を探索する場合には，実質的に Random Search と大差のない探索の挙動をすると考えられる．

図 3 に提案手法および auto-sklearn 実行時の交差検証精度の推移を示す．提案手法は Phase.2 でサンプルデータを用いて精度向上に有望なパイプラインを短時間で絞り込むことにより，Phase.3 において auto-sklearn よりも高精度な予測モデルを得ることができている．図 3(c) に auto-sklearn を 10 倍の時間実行したときの交差検証精度の推移を示す．auto-sklearn は探索初期に得た予測モデルから予測精度が向上しておらず，提案手法の予測精度に到達することはできていない．

これらの結果より，限られた時間内に機械学習パイプライン探索を行いたい場合には，提案手法のようにサンプルデータを用いた有望なパイプライン探索を予め行うことにより，効率的に高精度な予測モデルを構築することができると考えられる．

5. おわりに

本論文では，非データ分析専門家でもデータ分析に簡単に従事できることを目的として，サンプリングを用いた機械学習パイプライン探索方法を提案した．既存の機械学習の自動化を目的としたツール群は Random Search と大きく精度差が見られないという状況を踏まえ，データ分析専門家の作業工程に着想を得て，提案手法は前処理やハイパーパラメータの有効性を逐次的に確認しながら探索を進める．全てのデータを使った有効性確認は計算時間の観点で現実的ではないため，サンプルデータを用いて有望なパイプラインを探索した後に，全てのデータを用いて予測モデルを構築する．探索過程においてハイパーパラメータ探索空間の絞り込みやパイプライン選択を行うことにより，計算時間面での効率化を行った．

実験によって，提案手法は state-of-the-art な機械学習の自動化ツールの 1 つである auto-sklearn に対して，67% のデータにて高精度な予測モデルが構築可能であり，残りについても同等精度の予測モデルが構築可能であることを示した．

今後の課題として，回帰や推薦といった分類以外の機械学習タスクに提案手法の応用を広げることや，よりサイズの大きいデータセットに対する提案手法の評価が挙げられる．手法改善面では，3.4.3 節に示したパイプライン選択基準の詳細な性能評価や，各分類器の特性を踏まえた探索条件設定による効率化などが挙げられる．

参考文献

[1] IDC , Worldwide Big Data and Business Analytics Revenues Forecast to Reach $187 Billion in 2019, According to IDC , 2016, https://www.idc.com/getdoc.jsp?containerId=prUS41306516

[2] Chris Thornton, Frank Hutter, Holger Hoos, and Kevin Leyton-Brown, Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms, In Proc. of KDD 2013, 2013.

[3] Feurer, Matthias and Klein, Aaron and Eggensperger, Katharina and Springenberg, Jost and Blum, Manuel and Hutter, Frank, Efficient and Robust Automated Machine Learning, Advances in Neural Information Processing Systems 28, pp. 2944-2952, 2015.

[4] Isabelle Guyon, Imad Chaabane, Hugo Jair Escalante, Sergio Escalera, Damir Jajetic, James Robert Lloyd, Nuria Macia, Bisakha Ray, Lukasz Romaszko, Michele Sebag, Alexander Statnikov, Sebastien Treguer, Evelyne Viegas, A brief Review of the ChaLearn AutoML Challenge: Any-time Any-dataset Learning without Human Intervention, AutoML workshop, ICML, JMLR proceedings, 2016.

[5] Frank Hutter, Holger Hoos, and Kevin Leyton-Brown. Sequential Model-Based Optimization for General Algorithm Configuration In LION-5, 2011.

[6] Lisha Li, Kevin Jamieson, Giulia Desalvo, Afshin Rostamizadeh and Ameet Talwalkar. A Novel Bandit-Based Approach to Hyperparameter Optimization, AutoML workshop, ICML, JMLR proceedings, 2016.

[7] Guo-Xun Yuan, Chia-Hua Ho, Chih-Jen Lin: Recent Advances of Large-Scale Linear Classification. Proceedings of the IEEE 100(9): 2584-2603 (2012)

[8] Bergstra, James S., et al. “ Algorithms for hyper-parameter optimization. ” Advances in Neural Information Processing Systems. 2011.

[9] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/.