روش کشف کلاه برداری از طریق پردازش کلان داده در سیستم­های پرداخت اعتباری

محمد حسین مطیع بیرجندی

دانشکده مهندسی برق و کامپیوتر دانشگاه تهران [moti.hosein@ut.ac.ir](mailto:moti.hosein@ut.ac.ir)

# *چکیده:*

کارت­ها اعتباری و کارت­های پیش پرداخته[[1]](#footnote-1) بسیار محبوب و پر استفاده هستند. از این کارت­ها بسادگی می­توان جهت کلاهبرداری استفاده نمود و در زمینه کلاهبرداری بسیار آسیب پذیر هستند. هدف از این مقاله ارائه یک راه حل اتوماتیک و کارآمد جهت تشخیص این کلاهبرداری ها و تمرکز بر روی کاهش هشدار­هایی که در اثر تشخیص اشتباه کلاهبرداری به صدا درمی­آیند، می­باشد. بر خلاف الگوریتم­های موجود که براساس رفتار مشتری عمل می­کنند و نمی­توانند تا دنباله­ای از کلاهبرداری ها را تشخیص دهند و یا با تغییر رفتار مشتری دچار خطا می­شوند، تکنیک و الگوریتم پیشنهادی هر دو مشکل را حل کرده است. در انتها تکنیک ارائه شده بر روی مجموعه­ی واقعی از تراکنش های اعتباری در بازه سال 2012-2010 اعمال شده است. این روش می­تواند بصورت بی­درنگ تقلب را تشخیص دهد.

# مقدمه:

امروزه تراکنش­های بانکی زیادی در هر روز صورت می­پذیرد. با افزایش فروشگاه­های آنلاین و افزایش خرید از آنها و نیز توسعه تجارت الکترونیک در دنیای امروز، تقلب و کلاهبرداری در معاملات بانکی و تراکنش­های روزانه مشتریان بانک­ها در حال افزایش است. سالانه میلیون­ها دلار ضرر به صنعت بانکداری و مشتریان بانک­ها در اثر کلاهبرداری­های گوناگون وارد می­شود. بعنوان مثال در سال 2016 بیش از 16میلیارد دلار در ایالات متحده آمریکا به مشتریان بانکی ضرر وارد شد.[1] از سوی دیگر تعداد تراکنش­های کارت­های اعتباری در طول یک روز بسیار زیاد است، که این امر مستلزم استفاده از کلان­داده و ابزاری­های مدل سازی مرتبط با آن می­باشد. در زمان خرید، زمانی که یک تراکنش جعلی تشخیص داده شد سیستم تشخیص کلاهبرداری می­تواند بعنوان یک عامل پیشگیری کننده از انجام تراکنش ممانعت بعمل آورد. حتی تشخیص مبالغ پایین کلاهبرداری نیز اهمیت بالایی دارد زیرا که سارقین معمولا کارت­های مسروقی را با مبالغ پایین تست می­کنند.[2] روش­های تشخیص کلاهبرداری را می­توان به دو دسته اصلی تقسیم نمود: تشخیص سوءاستفاده و تشخیص ناهنجاری. در روش­های تشخیص سوءاستفاده سیستم براساس تراکنش­هایی که می­دانیم در آنها کلاهبرداری رخ داده­است، تمرین داده می­شود و تنها می­تواند کلاهبرداری­هایی را که مشابه آن در داده­هایی که روی آن آموزش داده شده است را تشخیص دهد. در روش­های تشخیص ناهنجاری سیستم روی تراکنش­های نرمال آموزش داده می­شود لذا توانایی کشف کلاهبرداری­ با روش­های جدید را داراست.[3] با افزایش روزافزون پیش بینی کننده­های در دسترس، روش­های یادگیری آماری که بطور موثر یک مدل پارسیونی[[2]](#footnote-2) با عمکرد پیش­بینی برتر معرفی می­کنند، یک روش ایده­آل برای توسعه سیستم­های تشخیص کلاهبرداری محسوب می­شوند. اما از سوی دیگر حجم بالای داده موجب می­شود تا پیاده سازی و توسعه ساده­ترین روش ها بصورت موثر کاری سخت یا حتی غیر ممکن گردد. اگرچه در نگاه اول بنظر می­رسد که سخت افزار­های جدید و تکنیک­های جدید مورد استفاده در کلان­داده مانند پردازش ابری، بسیاری از مشکلات را حل کرده­اند اما مشکلاتی مانند هزینه منابع مصرفی، سربار محاسباتی و مسئله نقض حریم خصوصی و امنیت کاربران باعث می­شوند تا عمده این روش­ها کمتر جذاب و قابل استفاده باشند. در عوض یک روش ساده­تر می­تواند مشکلات کلان داده را بصورت الگوریتمی حل نماید. روش­های گوناگون و مختلفی شامل نمونه برداری تصادفی، شکست و سرهم سازی و روش­های یادگیری آنلاین برای تحلیل کلان­داده توسعه داده شده است. درحالیکه برخی روش­ها برای ساده سازی انتخاب مدل توسعه یافته­اند، تمرکز در درجه اول بر روی تکنیک­های منظم سازی بوده است، که می­تواند در زمان کار با ساختمان داده­های پیچیده مانند مجموعه داده همبسته غیرگوسی سخت و چالش برانگیز باشد. بنابراین به تکنیک­های موثری نیاز داریم که بتوانند ساختمان­داده­های پیچیده مانند آنچه بیان شد را بصورت ساده­تر تحلیل و پردازش کنند.

"انتخاب مرحله­ای" یک روش انتخاب کلاسیک است که بعنوان یک جایگزین برای تکنیک­های منظم سازی معروفی مانند Lasso که مجددا مورد توجه محققین قرار گرفته است، استفاده می­شود. "انتخاب مرحله­ای" با اینکه ارتباط قوی با تکنیک منظم سازی دارد، انعطاف پذیری بیشتری برای تعامل با ساختمان داده­های پیچیده تر دارد. فرضیه پایه­ای روش "انتخاب مرحله­ای" اجرای روش "آهسته دمیدن" است. همانطور که از نام روش "انتخاب مرحله­ای" انتظار داریم، فرایند انتخاب مرحله­ای با یک مدل خالی شروع می­شود و در طی چندین مرحله تکرار یادگیری موثر، مدل ساخته می­گردد.

این پژوهش خانواده­ای جدید از روش­های مرحله­ای تصادفی را معرفی می­کند که از زیرنمونه برداری[[3]](#footnote-3) جهت حل مسئله انتخاب مدل در کلان­داده با قابلیت پشتیبانی از ساختمان داده­های پیچیده استفاده می­کند. در ادامه ابتدا به بررسی مسائل تئوری پیرامون انتخاب مرحله­ای می­پردازیم. لذا در ابتدا روش انتخاب مرحله­ای تصادفی را بررسی می­کنیم. سپس به بررسی داده­های خوشه­ای می­پردازیم. در انتهای این بخش به بررسی معادله تخمینی انتخاب مرحله­ای تصادفی خواهیم پرداخت. با پایان این بخش به شبیه سازی داده­های گوسی و باینری می­پردازیم. در بخش بعد به بررسی تشخیص کلاهبرداری به کمک روش­های بیان شده خواهیم پرداخت. در نهایت در بخش جمع بندی به بررسی و مقایسه نتایج روش­های معرفی شده روی دو نوع داده­ای که در بخش شبیه سازی بررسی شد، می­پردازیم.

# روش­ها

## روش زیر نمونه برداری (Sub-sampling)

روش های زیرنمونه­برداری بر روی ایده انتخاب تصادفی داده جهت افزایش کارایی تمرکز دارند. ساده ترین این روش­ها، انتخاب تصادفی داده از مجموعه داده اولیه و انجام کار­های دلخواه با آن می­باشد. با توجه به ذات احتمالاتی این روش­ها، معمولا چندین بار تکرار این فرآیند و بررسی نتایج بدست­آمده با bootstrap های محبوب و معروف، بسیار رایج است. به تازگی توجه زیادی به این روش­های تصادفی شده است. بعنوان نمونه یک افزونه برای روش رگرسیون زیرنمونه­برداری کلاسیک منتشر کرده­اند که "بهره­برداری" نامیده می­شود (Leveraging for big data regression). در این روش توزیع مشاهدات یکنواخت نیست بلکه، یک مکانیزم مانند بهره­برداری (leveraging) برای اطمینان از اینکه برخی مشاهدات بیشتر از برخی دیگر نمونه برداری شده­اند، استفاده می­شود. G. Vaughan در [4] بیان کرده است که [5] می­گوید که زیرنمونه­هایی که در فرآیند زیرنمونه­برداری یا بهره­برداری بدست آمده­اند، می­توانند بعنوان جایگزین برای مصور کردن تمام مجموعه داده مورد استفاده قرار بگیرند.

برخی کار­ها در کاوش بدنبال کارکردهای بالقوه برای روش­های تصادفی در روش­های یادگیری آماری انجام شده است. مثلا در [4] آمده است که آقای Ahmed et al. در [6] انتخاب کردن به روش زیرنمونه­برداری/پایداری را معرفی کرده است که، عطف به روش رگرسیون لاسو که برای مشخص کردن عدم توازن کلاس­ها در مدل انتخابی است، می­باشد. براساس دانسته های نویسنده، روش­های تصادفی برای انتخاب بصورت مرحله­ای بررسی نشده­اند.

## روش مرحله­ای

روش انتخاب مرحله­ای یک روش کلاسیک انتخاب مدل است که به تازگی توجه بسیاری را به خود جلب کرده­است. ساختار پایه­ای تمام روش­های مرحلهای بدین صورت است که با یک مدل تهی (خالی) که با بردار ضرایب توصیف شده است، آغاز می­شوند. سپس پس از چندین بار پیمایش بردار ضرایب براساس گام­های ساده­ی یادگیری که محاسبه شده اند، بروزرسانی می­شود. یک مثال برای تکنیک مرحله­ای، روش انتخاب مرحله­ای رو به جلو (forward stagewise selection) است. در این روش از یک مدل خطی چند متغییره استفاده می­شود. در پیمایش t ام در این روش، الگوریتم i را اینگونه تعریف می­کند: تعریف می­کند که یک همبستگی میان j امین covariate و باقی­مانده مدل براساس مقادیر فعلی ضرایب است. سپس بردار ضرایب مطابق فرمول زیر بروزرسانی می­گردد.

منظور از در این فرمول، سایز ثابت گام یادگیری می­باشد که می­بایست به اندازه کافی کوچک باشد و برای تمام داریم . دنباله یا "مسیر" (path) بدست آمده از تقریب ضرایب می­تواند مشابه روش­های تنظیمی (regularization) برای انجام انتخاب مدل مورد استفاده قرار گیرد.

میان روش­های مرحلهای و روش­های تنظیم(regularization) ملاحظه شده یک ارتباط قوی برقرار است. بعنوان مثال، نشان داده شده است که مسیری که توسط روش انتخاب مرحله­ای رو به جلو تولید می­شود به شرطی که برقرار باشد، به مسیری که توسط lasso تولید می­شود میل می­کند. بعلاوه ثابت شده است که دسته کلی تکنیک­های مرحله­ای با تقریب خوبی با همتایان خود در تکنیک­های regularization برابری می­کنند. حتی در صورتی که این ارتباط قوی میان تکنیک­های مرحله­ای و تکنیک­های regularization را کنار بگذاریم، روش­های مرحله­ای هم در کاربری­های مختلف منعطف هستند و هم از نظر محاسبات کامپیوتری بهینه هستند.

## روش انتخاب مرحله­ای تصادفی

ما با مشاهده کاربرد­های زیرنمونه­برداری و ترکیب آن با روش مرحله­ای، روش انتخاب مرحلهای تصادفی را مطرح می­کنیم. مدل را اینگونه بیان می­کنیم: پاسخ ما و برای covariates مشاهدات می­باشد که است. تابع ضرر (loss function) از فرمی نامعین، بدون شکستگی و محدب است. برای می­توان گردایان را محاسبه کرد و نامعین نیست. احتمال نمونه برداری برای هر مشاهده باشد و نشان دهنده یک نمونه رندم با سایز از عداد طبیعی از 1 تا که نسبت زیرنمونه­برداری است که در بازه 0 تا 1 قرار دارد. در صورتی که باشد، آنگاه روش مرحله­ای بصورت کلاسیک و قابل پیش بینی خواهد بود. ما نمونه تصادفی تمام مجموعه داده را در گام ام بصورت نمایش میدهیم. با شروع از برای داریم:

که تابع ضرری است که در تنها با استفاده از محاسبه شده است. برای سادگی ما از نرم نوع 1 برای استفاده کرده­ایم که در فرمت بروزرسانی که در تحقیقات گذشته بخوبی مطالعه شده است، تاثیر گذار است. اما با این حال همچنان می­توان از خیلی از توابع پنالتی محدب استفاده نمود.

انتخاب ها از دنباله احتمالات و نمونه برداری با جایگزینی یا بدون آن می­تواند براساس مسئله ما باشند و حکم کلی ندارند. در [4] برای قسمت زیرنمونه­برداری در روش stochastic gradient boosting آقای Friedman در[7] از یک توزیع یکنواخت و نمونه­برداری بدون جایگزینی استفاده کرده است. از سوی دیگر در [5] یک توزیع از زیرنمونه­ها برای روش leverage linear regression با استفاده از قدرت مشاهدات ایجاد کرد تا مقادیر احتمالات را دیکته کرده و نمونه­برداری با جایگزینی انجام دهد. بعلاوه می­توانیم توابع توزیع تطبیقی را براساس کارآیی مدل مانند قدرمطلق خطا (absolute error) یا misclassification که جایگزین توزیع نمونه­برداری می­شوند تا بر روی مشاهداتی که پیش بینی آن ها سخت است تمرکز کند، را تصور کنیم. فرکانس نمونه برداری نیز می­تواند تنظیم شود. مثلا ممکن است یک زیرنمونه قبل از پیمایش اول مورد استفاده مجدد قرار بگیرد و یا زیرنمونه­های جدید می توانند در بازه­های مشخص برداشته شوند.

در عمل اندازه گام یا و مکانیزم پایان الگوریتم باید با دقت مورد توجه قرار گیرد. ارتباط میان الگوریتم­های مرحله­ای و regularization به سایز کوچک گام ها تکیه دارد. از طرفی سایز کوچک گام­ها موجب اتلاف محاسبات و منابع محاسباتی می­شود. استفاده از مجموعه ساده­ای از قوانین پایان که براساس تعداد ویژگی­ها و تعداد پیمایش ها هستند بطور کلی توصیه می­شوند. اول، دانستن پیش بینی اینکه چه تعدادی از ویژگی­های ممکن است بر روی پاسخ تاثیر دارند، می­تواند به دانستن تعداد پیش بینی کننده­های مورد استفاده در مدل کمک کند. سپس حداکثر تعداد پیمایش­ها باید مشخص شود.

## داده­های خوشه­ای

بعنوان مشکل انگیزشی در این پژوهش، داده ها به فرمت خوشه­ای هستند که ما انتظار داریم مشاهدات برای یک دارنده کارت بشدت همبستگی داشته باشند. استفاده از این همبستگی می­تواند بهره­ وری روش ما را بهبود بخشد. یک روش برای کار کردن با دادههای همبسته غیر گوسی استفاده از Generalized Estimation Equation یا GEE هاست. می­توان به GEE ها بعنوان نسخه جامعیت داده شده از Generalized Linear Model یا GLM ها نگریست، که از آن بطور مستقیم جهت مدل سازی داده­های خوشه بندی شده که این باور وجود دارد که میان خوشه­ها یک همبستگی وجود دارد. تفاوت اصلی این است که برخلاف GLM، در GEE فرض نمی­شود یک likelihood کامل یا یک ساختار توزیعی داریم. در عوض فقط یک ساختار متوسط حاشیه­ای و واریانس بر اساس یک پیش بینی کننده خطی و یک پارامتر پراکنش مزاحم فرض می شود که وجود دارد.

## معادله تخمینی انتخاب مرحله­ای

ترکیب کردن GEE ها در فریمورک تصادفی مرحله­ای در ابتدا نیاز دارد تا معادلات تخمینی را برای گرادیان­های نامعین جایگزین کنیم. سپس تغییرات کمی اعمال می­گردد تا خوشه بندی در نظر گرفته شود و بازدهی در روش معادلات تخمینی انتخاب تصادفی مرحله­ای بدست آید. بطور مشابه برای روش تصادفی مرحله­ای داریم: احتمالات نمونه­برداری برای هر خوشه به ازای و نشان دهنده یک نمونه رندم با سایز از عداد طبیعی از 1 تا است که نسبت زیرنمونه­برداری است که در بازه 0 تا 1 قرار دارد. نمونه تصادفی کل مجموعه داده در مرحله بصورت داده شده است و معادله تخمینی که براساس این نمونه از داده در و محاسبه شده است بصورت نمایش داده می­شود. با شروع از برای داریم:

که

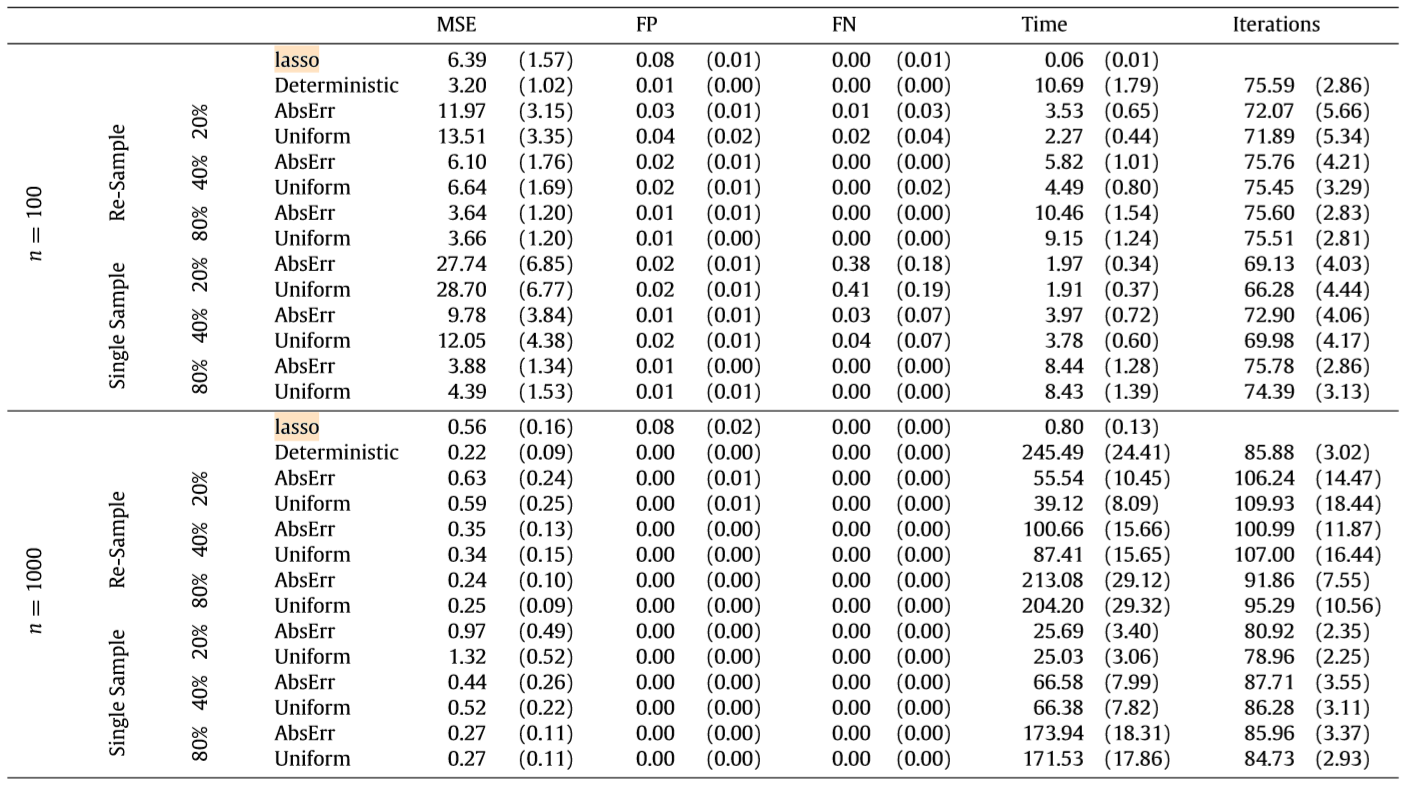
مشابه فرم کلی روش تصادفی مرحله­ای، تعداد زیادی تابع پنالتی وجود دارند که می­توانند مورد استفاده قرار بگیرند اما ما تنها بر روی نرم 1 یعنی تمرکز می­کنیم.

مزیت شاخص و یکتا در کار کردن با داده های خوشه بندی شده این است که بجای اینکه مشاهده­های جداگانه را نمونه برداری کنیم، براساس خوشه نمونه برداری انجام می­دهیم. در نمونه برداری خوشه­ای مدت زمان محاسبه کاهش می­یابد و استفاده از منابع محاسباتی کاهش می­یابد. علت این امر آن است که در هر گام تنها تعداد ثابت و کمی از خوشه­ها باید در نظر گرفته شوند.

# شبیه­سازی

## داده­های گوسی

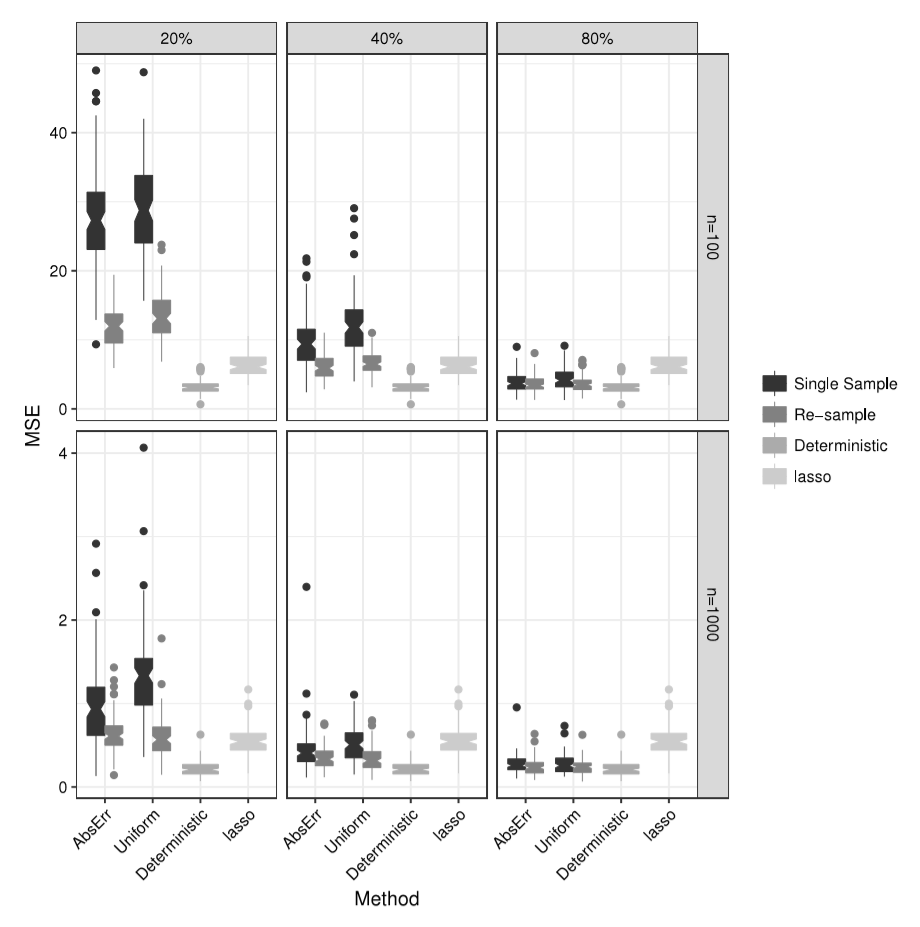
ما فرایند شبیه سازی را در جهت طولی و با سایز خوشه با انجام دادیم. ما دو مجموعه داده با سایز های مختلف را که خوشه و خوشه آزمایش کردیم. ما در ابتدا حالتی را در نظر می­گیریم که پاسخ از توزیع گوسی روی covariate ها پیروی می­کند. تمام covariate ها بصورت مستقل از روی توزیع نرمال استاندارد تولید شده­اند. از 500 تا از ضرایب، 10 تا از توزیع یکنواخت در بازه 1.5 تا 2.5 نمونه­برداری شده بودند و مابقی صفر بودند. هر بردار پاسخ با فرمت برای هر خوشه از توزیع نرمال چند متغییره با میانگین و مقدار و ماتریس کوواریانس که دارای المان­های مورب است و همبتسگی میان جفت خوشه ها با مقدار 0.6 ، تولید شده است.

**

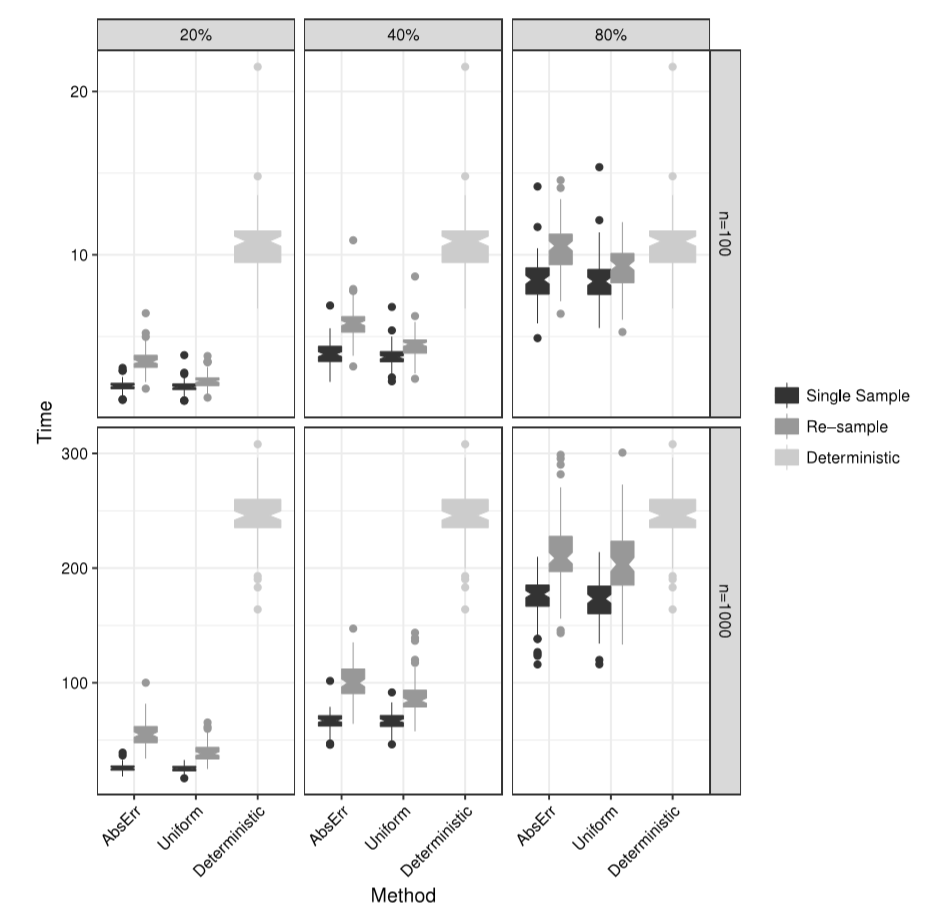
جدول 1: نتایج شبیه سازی داده­های گوسی برای تعداد 100 و 1000 خوشه

جدول 1 گزارش نتایج شبیه سازی در محیط و پارامتر­هایی است که بیان شد. در این شبیه سازی تمام تکنیک های موجود اجرا و تست شده اند و در جدول 1 نتایج­شان در کنار یکدیگر آمده است.

در تصویر 1 نمودار جعبه­ای اندازه­گیری های پیشگویانه قابل ملاحظه است. در این نمودار زمان اجرای تمام تکنیک­ها با مرجعیت روش­های قابل پیش بینی و لاسو رسم گردیده است. پیاده سازی لاسو توسط glmnet انجام شده است و بیشتر محاسباتش را از طریق زبان Fortran انجام می­دهد. درحالیکه تکنیک های بیان شده، با زبان R پیاده سازی شده­اند که به مراتب از زبان­های کامپایلری کند تر هستند و Fortran یک زبان کامپایلری است. بعلاوه لاسو در حالت کلی فرض می­کند که مشاهدات مستقل هستند، که این فرض موجب ساده تر شدن پیچیدگی محاسبات نسبت به حالتی که فرض شود یک همبستگی میان خوشه­ها وجود دارد. این تفاوت ها موجب تفاوت زمانی غیر قابل مقایسه­ای میان روش لاسو و روش­های بیان شده در این مقاله شده است که این به این علت در تصویر 2 روش لاسو حذف شده است.



تصویر 1: نمودار جعبه­ای اندازه­گیری­های پیشگویانه روی 100 خوشه



تصویر 2: نمودار جعبه­ای زمان اجرای الگوریتم­ها برحسب ثانیه بر روی 100 خوشه

# جمع بندی

حجم بسرعت در حال افزایش و پیچیدگی داده نشان می­دهد که چگونه روش­های قدیمی پیش بینی دوباره بازآفرینی می­شوند و بهبود می­یابند تا بتوانند با چالش های جدید مواجه شوند. با اینکه روش­های سنتی regularization گام­های عالی برداشته بودند اما برخی داده ساختار­ها مانند longitudinal همچنان ایجاد چالش می­کنند. بعلاوه اعمال مستقیم این روش­ها غیر قابل قیاس پذیری است، بنابراین یک زیرنمونه برداری ساده می­تواند اجرا شود تا یک مجموعه داده بزرگ را مدیریت کند، اما انجام این کار بسیاری از اطلاعات را بدون استفاده می­گذارد. روش انتخاب مرحله­ای تصادفی می­تواند یک زیرنمونه­برداری جدید در دو جهت را معرفی کند، تا منابع محاسباتی را حفظ کند و خطای پیش بینی را کمینه کند. یک زیرنمونه می­تواند انتخاب شود و روش مرحله­ای روی تنها آن زیرنمونه اعمال گردد.

محدودیت­های کار ارائه شده در این مقاله مسیرهای جذابی را برای مطالعات آینده نمایش می­دهد. روش انتخاب مرحله­ای تصادفی می­تواند از هر نوع تابع پنالتی محدبی استفاده کند اما اینجا تنها از نرم 1 استفاده شده است. پیاده سازی این روش ها با کمک سایر این توابع می­تواند اجازه­ی استفاده از مدل­های پیچیده تر در آینده را بدهد.

# منابع:

[1] M. Kozubovska, “Breaking up big banks,” *Research in International Business and Finance*, vol. 41, pp. 198–219, Oct. 2017.

[2] W. N. Robinson and A. Aria, “Sequential fraud detection for prepaid cards using hidden Markov model divergence,” *Expert Systems with Applications*, vol. 91, pp. 235–251, Jan. 2018.

[3] L. Seyedhossein and M. R. Hashemi, “Mining information from credit card time series for timelier fraud detection,” in *2010 5th International Symposium on Telecommunications*, 2010, pp. 619–624.

[4] G. Vaughan, “Efficient big data model selection with applications to fraud detection,” *International Journal of Forecasting*, Jun. 2018.

[5] P. Ma and X. Sun, “Leveraging for big data regression,” *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 7, no. 1, pp. 70–76, Jan. 2015.

[6] I. Ahmed, A. Pariente, and P. Tubert-Bitter, “Class-imbalanced subsampling lasso algorithm for discovering adverse drug reactions,” *Statistical Methods in Medical Research*, vol. 27, no. 3, pp. 785–797, Mar. 2018.

[7] J. H. Friedman, “Stochastic gradient boosting,” *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, Feb. 2002.

1. Prepaid cards [↑](#footnote-ref-1)
2. Parsimonious model [↑](#footnote-ref-2)
3. Sub-sampling [↑](#footnote-ref-3)