dataacademy.ir

پیش‌بینی ریزش مشتری با استفاده از تکنیک‌های داده کاوی: مبتنی بر ماشین بردار و الگوریتم ژنتیک

سید محمد سید حسینی، بهروز مینایی، مریم قدمی*، امیرحسین زمانیان

دکترای مهندسی صنایع، استاد دانشکده مهندسی صنایع، دانشگاه علم و صنعت ایران، seyedhosseini @ iust.ac.ir

دکترای مهندسی و علوم کامپیوتر، استادیار دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، b_minaei@iust.ac.ir

دانشجوی کارشناسی ارشد مهندسی صنایع، دانشگاه علم و صنعت ایران، ghadami_maryam@ind.iust.ac.ir

کارشناسی ارشد مهندسی مکانیک، دانشگاه امیرکبیر، zamanian.amir@gmail.com

* نویسنده عهده‌دار مکاتبات، ghadami_maryam@ind.iust.ac.ir

نشانی پستی: تهران، نارمک، خیابان آیت، خیابان چمن شرقی، پلاک ۴۴، طبقه پنجم.

شماره تلفن: ۰۲۱۷۷۹۰۴۶۷۰- ۰۹۱۲۳۷۲۰۸۷۸

پیش‌بینی ریزش مشتری با استفاده از تکنیک‌های داده کاوی: مبتنی بر ماشین بردار و الگوریتم ژنتیک

چکیده

با افزایش رقابت در بازارهای جهانی، حفظ مشتری به عنوان یکی از مهم‌ترین مسائل در شرکت‌ها مطرح شده است. به گونه‌ای که پیشگیری از ریزش مشتری، بخش مهمی از مدیریت ارتباط با مشتری[2]، تلقی می‌گردد. در این میان، چگونگی پیش‌بینی و پیشگیری از ریزش مشتریان، توجه بسیاری از شرکت‌ها و نیز پژوهشگران این حوزه را به خود جلب نموده است. از سوی دیگر، با خودکارسازی جریان عملیات، مجموعه‌های عظیمی از داده‌های مربوط به عملیات‌های روزانه جمع‌آوری می‌گردد و بستر مناسبی جهت بهره‌گیری از تکنیک‌های داده‌کاوی فراهم می‌آورد. ماشین بردار پشتیبان، ابزاری قدرتمند در طبقه‌بندی به شمار می‌رود و می‌تواند مسائلی را حل نماید که رویکردهای سنتی گذشته در حل آن‌ها ناتوان هستند. در این تحقیق از یک مدل مبتنی بر الگوریتم ژنتیک و ماشین بردار پشتیبان به منظور پیش‌بینی ریزش مشتریان یک فروشگاه بزرگ تامین‌کننده کالا و اقلام مختلف بهره گرفته شده است. در مرحله اول در این مدل از الگوریتم ژنتیک به منظور انتخاب ویژگی و تعیین مقادیر بهینه پارامترهای ماشین بردار پشتیبان، به صورت همزمان، استفاده می‌شود و سپس مدل پیش‌بینی با توجه به مقادیر پارامترهای تعیین شده ساخته می‌شود. عملکرد مدل با روشهای رده‌بندی دیگر همچون درخت تصمیم، استدلال مبتنی بر مورد و شبکه عصبی مصنوعی مورد مقایسه قرار گرفته است و نتایج نشان می‌دهد به کارگیری مدل مذکور در پیش‌بینی ریزش مشتری و شناسایی عوامل موثر در حفظ مشتری، به میزان قابل توجهی موثر بوده است.

کلمات کلیدی: پیش‌بینی ریزش مشتری، ماشین بردار پشتیبان (SVM)[3]، الگوریتم ژنتیک، داده‌کاوی.

dataacademy.ir

[2] Customer Relationship Management

[3] Support Vector Machine

Churn customer prediction using data mining: based on support vector machine and gentic algorithm

M.Hoseini, B. Minaei, M.ghadami, A. Zamanian
Industrial Engineering Department, Iran University of Science and Technology, seyedhosseini @ iust.ac.ir
Computer Engineering Department, Iran University of Science and Technology, b_minaei@iust.ac.ir
Industrial Engineering Department, Iran University of Science and Technology, ghadami_maryam@ind.iust.ac.ir
Machanical Engineering Department, Amirkabir University of Technology , zamanian.amir@gmail.com

Abstract

As markets become increasingly saturated, companies have acknowledged that their business strategies should focus on identifying those customers who are likely to churn. Customer churns analysis and predication is an important part of Customer Relationship Management (CRM). With the purpose of retaining customers, academic as well as practitioners find it crucial to build a churn prediction model that is as accurate as possible. In other side, the development of automated data collection tools and the imperative need for the interpretation and exploitation of massive data volumes have resulted the development and flourishing of data mining techniques. Support Vector Machines (SVM) algorithm is one of the most effective machine learning algorithms and successfully solves classification problems in many domains. In this paper a hybrid model, combining genetic algorithms with SVM classifier, is employed to predict customer churn in Mondrian food mart department store. The procedure includes two steps. In the first step genetic algorithm determines SVM parameter values while discovering a subset of features and in second step customer churn prediction model is established based on determined parameters. Empirical results are compared with some classifiers such as simple SVM, Decision tree, Case Based Reasoning and Artificial Neural Networks, and indicate that the predictive performance of hybrid model in specific case is more effective.

dataacademy.ir

Keywords

Churn customer prediction, Support Vector Machine (SVM), Genetic algorithm, data mining.

۱- مقدمه

همزمان با اشباع بازار و افزایش رقابت جهت ادامه حیات و حفظ موقعیت تجاری، سازمانها استراتژیهای کسب و کار خود را به سوی شناسایی مشتریان احتمالی که سازمان را ترک خواهند نمود و ارائه راهکارهای مناسب جهت پیشگیری از آن متمرکز نمودند. اصطلاح ریزش مشتری[۴]، به ترک مشتری و دریافت کالا و یا خدمات از سایر شرکت‌های رقیب تلقی می-گردد و معمولاً در غالب مسائل رده‌بندی[۵]، با تقسیم مشتریان به دو دسته مشتریانی که سازمان را ترک می‌کنند و مشتریانی که به سازمان وفادار خواهند ماند، مورد بررسی قرار می‌گیرد(کانگ و پی‌جی[۱]، ۲۰۰۸). پیش‌بینی ریزش مشتری، برای اولین‌بار در صنعت ارتباطات بی‌سیم و به منظور تشریح زیان و خساراتی که متوجه مشتریان و متقاضیان می‌گشت، مطرح گردید. همزمان با اشباع بازار و تشدید رقابت، اهمیت ریزش مشتری در بسیاری از صنایع همچون شرکت‌های بیمه، سرمایه‌گذاری محصولات، ارائه‌دهندگان خدمات اینترنتی و ارتباطاتی، بانکداری، ارائه‌دهندگان خدمات بهداشتی و ... چندین برابر گردید(اکسین و همکاران[۲]، ۲۰۰۹). پیشگیری از ریزش مشتری به عنوان بخش مهمی از مدیریت ارتباط با مشتری، تاثیر به سزایی بر افزایش سهم بازار، کاهش هزینه‌ها و سایر ابزارهای رقابتی دارد. از سوی دیگر هزینه حفظ مشتری موجود بسیار کم‌تر از هزینه صرف شده جهت جذب یک مشتری جدید می‌باشد. طبق برآورد انجام شده (اکسین و همکاران[۲]، ۲۰۰۹) ، ۱٪ کاهش نرخ ریزش مشتری منجر به ۶٪ افزایش در میزان سود شرکت می‌گردد و این مسئله اهمیت قابل توجهی در سودآوری و رشد شرکت‌ها دارد. به منظور پیشگیری از ریزش مشتری، رفتار مشتری با توجه به مجموعه اطلاعات و داده‌های جمع‌آوری شده تحلیل، مشتریان مستعد ترک (ریزش) شناسایی و درنهایت با هدف قراردادن این دسته از مشتریان، استراتژیهای مناسب و موثر جهت حفظ آنها طرح‌ریزی و اجرا می‌گردد. در این راستا، ایجاد یک مدل دقیق و کارا نقش تعیین‌کننده‌ای در مدیریت ریزش مشتری دارد(کوزمنت و واندن[۳]، ۲۰۰۸). تحقیقات وسیعی در حوزه پیش‌بینی ریزش مشتری صورت گرفته است و تکنیکهای داده‌کاوی زیادی همچون شبکه‌های بیزین، k نزدیکترین همسایگی، رگرسیون لجستیک، درخت تصمیم شبکه‌های عصبی مصنوعی، ماشین بردار پشتیبان و . . . جهت مدل‌سازی پیش‌بینی ارائه شده است که می‌تواند به طور قابل توجهی به شناسایی مشتریان مستعد ترک کمک نماید (هوانگ و همکاران[۴]، ۲۰۱۰). «بین و همکاران» با استفاده از تکنیک درخت تصمیم، یک مدل رده‌بندی باینری جهت پیش‌بینی ریزش مشتری ارائه دادند (بین و همکاران[۵]، ۲۰۰۷). «ایکسو و همکاران» از شبکه‌های عصبی مصنوعی با انتشار خطا به عقب (BPN) برای ساخت مدل پیش‌بینی استفاده کردند(ایکسو و همکاران[۶]، ۲۰۰۶). در مدل ارائه شده به منظور بهبود کیفیت مجموعه داده (کاهش اثرات منفی داده های غیردقیق، ناقص و

---

[4] Customer churn

[5] classification

متناقض بر عملکرد مدل) از rough set theory بهره گرفته شده است. «مینگ و همکاران» از شبکه‌های بیزین برای حل این مسائل استفاده کرده است (مینگ و همکاران[۷]، ۲۰۰۵). «پندهارکر» از یک مدل بهبود یافته شبکه‌های عصبی با استفاده از الگوریتم ژنتیک جهت پیش‌بینی ریزش مشتری در یک شرکت ارتباطات سیار استفاده نمود. تابع برازندگی در الگوریتم ژنتیک بر دو اساس حداکثر نمودن دقت مدل و مینیمم نمودن آنتروپی متقابل[۶] در نظر گرفته شده و در نهایت نتایج با یکدیگر مورد مقایسه قرار گرفته است (پندهارکر[۸]، ۲۰۰۹). مرور کلی بر کاربرد تکنیکهای داده‌کاوی در حوزه پیش‌بینی ریزش مشتری و رده‌بندی آنها توسط «خاک‌آبی و همکاران» صورت گرفته است (خاک‌آبی و همکاران[۹]، ۲۰۱۰).

با توجه به عملکرد مناسب و دقت بالای پیش‌بینی ماشین بردار پشتیبان در محدوده وسیعی از حوزه‌ها همچون بیوانفورماتیک (چن و همکاران[۱۰]، ۲۰۰۵)، سیستم ارزیابی اعطای وام به مشتری(لی و همکاران[۱۱]، ۲۰۰۶)، تخمین ارزش محصول (چن و وانگ[۱۲]، ۲۰۰۷)، شناسایی تغییرات شدت صوت(اسیر[۱۳]، ۲۰۰۶) ، رده‌بندی متون و تصاویر(براتکو و فیلیپیک[۱۴]، ۲۰۰۶، کیم و همکارن[۱۵]، 2005)، تصدیق خودکار چهره(بیسکو و همکاران[۱۶]، ۲۰۰۵) ، تشخیص روش معالجه (گلاتسوس و همکاران[۱۷]، ۲۰۰۵) و ...، مشاهده می‌شود که در حوزه بازاریابی و پیش‌بینی ریزش مشتری مطالعات کمتری صورت گرفته شده است(کوزمنت و واندن[3]، ۲۰۰۸). «کیم و یون» از روش logit به‌منظور شناسایی عوامل تعیین‌کننده در ریزش و نیز وفاداری مشتریان در صنعت تلفن همراه استفاده نمودند(کیم و یون[۱۸]، ۲۰۰۴). «ژائو و لی» ماشین بردار پشتیبان را به دلیل برخورداری از دقت بالا و قدرت تعمیم‌پذیری مناسب، به عنوان مدل پیش‌بینی ریزش مشتری انتخاب نمودند و با مقایسه نتایج پیاده‌سازی مدل، بر مجموعه داده جمع‌آوری شده در صنعت مخابراتی بی‌سیم، با روشهای شبکه عصبی، درخت تصمیم و Naïve Bays، نشان دادند که ماشین بردار پشتیبان عملکرد بهتری داشته است(ژائو و لی[۱۹]، ۲۰۰۵). «کوزمنت و واندن» به منظور پیش‌بینی ریزش مشتری با تعداد نمونه سنگین، از حالت غیرخطی ماشین بردار پشتیبان (Kernel RBF) استفاده کردند(کوزمنت و واندن[3]، ۲۰۰۸). تخمین مقادیر بهینه پارامترهای مسئله، به‌کمک متد grid search انجام شده است. همچنین جهت بررسی اعتبارسنجی مدل، نتایج با دو روش رگرسیون لجستیک و random forests مورد مقایسه قرار گرفت. «جینگ و اکسینگ» با استفاده از ماشین بردار پشتیبان غیرخطی، مدلی جهت پیش‌بینی ریزش مشتریان بانک ارائه نمودند. در این تحقیق دقت مدل تحت توابع کرنل مختلف محاسبه و مورد مقایسه قرار گرفته و در نهایت تابع کرنل Cauchy به دلیل برخورداری از دقت بالاتر در پیش-بینی مجموعه داده مورد بررسی انتخاب می‌گردد(جینگ و اکسینگ[۲۰]، ۲۰۰۸). «اکسین و هونگ‌وانگ» مدلی مبتنی بر ماشین بردار پشتیبان ارائه نمودند که از تحلیل PCA در مرحله انتخاب ویژگی استفاده می نمود(اکسین و همکاران[2]، ۲۰۰۹). در

---

[6] cross entropy

حالیکه «کانگ و پی‌جی» از RFE برای انتخاب زیرمجموعه مناسب از ویژگیها و حذف ویژگی‌های غیرمرتبط بهره گرفتند(کانگ و پی‌جی[1]، ۲۰۰۸). همچنین «وانگ و نیو» به منظور پیش‌بینی ریزش مشتریان کارتهای اعتباری از مدلی مبتنی بر ماشین بردار پشتیبان استفاده نمودند که در آن انتخاب ویژگی به کمک Rough Set Theory تعیین می‌گردد(وانگ و نیو[21]، ۲۰۰۹). همان طور که مشاهده می‌شود انتخاب زیرمجموعه مناسبی از ویژگی‌ها به منظور رده‌بندی، مسئله مهمی در طراحی یک مدل محسوب می‌گردد. رویکرد متداول بدین گونه است که در فاز یادگیری و ساخت مدل، تا آن جایی که امکان دارد، ویژگی‌ها و مشخصه‌های بسیاری جمع‌آوری می‌شود. سپس تعیین بهینه زیرمجموعه کوچکتری از ویژگی‌ها برای طبقه‌بندی داده‌ها ضروری است و در صورت عدم انجام آن، کارایی مدل رده‌بندی، کاهش می‌یابد(لی و همکاران[22]، ۲۰۰۸).

در این مطالعه با استفاده از الگوریتم ژنتیک، انتخاب ویژگی و مقادیر مناسب پارامترهای مدل به صورت همزمان تعیین شده و در ادامه از مدل ماشین بردار پشتیبان، جهت رده‌بندی و پیش‌بینی عملکرد استفاده می‌گردد. در بخش دوم این مقاله، مقدمه‌ای بر ماشین بردار پشتیبان و الگوریتم ژنتیک بیان گردیده است. در بخش سوم نحوه عملکرد الگوریتم ترکیبی GA-SVM شرح داده می‌شود. نتایج تجربی پیاده‌سازی مدل بر روی مجموعه داده مشتریان یک فروشگاه بزرگ تامین‌کننده انواع مختلف کالا اعم از مواد غذایی، بهداشتی، لوازم خانگی و ... [7]در بخش چهارم مورد مقایسه قرار می‌گیرد و در انتها بخش پنجم به جمع‌بندی نتایج، اختصاص یافته است.

۲- ماشین بردار پشتیبان و الگوریتم ژنتیک

در این بخش مقدمه‌ای بر روش ماشین بردار پشتیبان، به عنوان یکی از روشهای دقیق در مسائل رده‌بندی، و الگوریتم ژنتیک به منظور تعیین زیرمجموعه‌ای مناسب از ویژگیها بیان می‌گردد.

۲-۱- ماشین بردار پشتیبان (SVM)

مسئله رده‌بندی، یکی از مسائل اصلی مطرح شده در یادگیری ماشین[8] است و در این میان، یکی از روش‌هایی که به صورت گسترده‌ای برای مسئله رده‌بندی مورد استفاده قرار می‌گیرد، روش ماشین بردار پشتیبان است. پایه‌گذاری ماشین بردار پشتیبان توسط «واپنیک» (واپنیک[23]، ۱۹۹۵) و بر اساس تئوری آموزش آماری پایه‌ریزی شده است و به دلیل خواص ویژه و عملکرد تجربی مطلوب در رده‌بندی، رایج گردید. فرمولاسیون ماشین بردار پشتیبان بر مبنای کمینه‌سازی ریسک ساختاری

---

[7] Mondrian food mart department store
[8] Machine Learning

(SRM)[9]استوار است و نشان داده شده است که از کمینه‌سازی ریسک تجربی (ERM)[10] که در شبکه‌های عصبی مصنوعی به کار گرفته می‌شود، بسیار بهتر عمل می‌نماید(واپنیک[23]، ۱۹۹۵). با توجه به گستردگی مفاهیم ماشین بردار پشتیبان، مقدمه‌ای از آن در این بخش آورده شده است و برای جزئیات بیشتر به مراجع (برگز[24]، ۱۹۹۸، گان[25]، ۱۹۹۸) مراجعه شود.

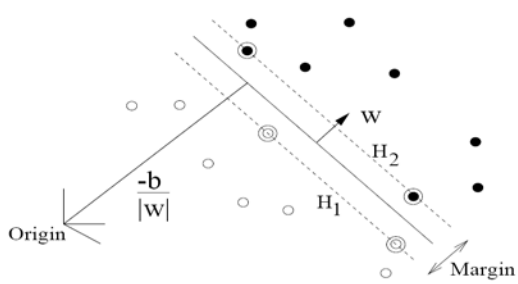
بدون از دست رفتن کلیت، مساله رده‌بندی می‌تواند به رده‌بندی بین دو رده محدود گردد (شکل۱). ماشین بردار پشتیبان را می‌توان پدید آورنده خط یا ابر صفحه‌ای در بین مجموعه داده‌ها برای رده‌بندی آنها در نظر گرفت. مساله‌ی جداسازی را در نظر بگیرید که در آن مجموعه بردارهای آموزش به دو دسته مجزا متعلق باشند.

$$\{(x_1,y_1)\ldots(x_m,y_m)\} \in X\times\{\pm 1\} \qquad (۱)$$

که با ابر صفحه‌ی معادله (۲) جدا می‌گردد.

$$\mathbf{w}.\mathbf{x}+b=0. \qquad (۲)$$

در حالت دو بعدی، عملکرد ماشین بردار پشتیبان می‌تواند به سهولت شرح داده شود، در این حالت SVM سعی در پیداکردن خطی دارد که دو دسته داده را توسط یک خط از هم جدا سازد. ممکن است خطوط بسیاری دارای این خاصیت باشند اما SVM سعی می‌نماید که خطی را در نظر بگیرد که بیشترین فاصله اقلیدسی بین نزدیک‌ترین داده به خط در هر دو دسته را داشته باشد. در این حالت، داده‌هایی که کمترین فاصله را با خط داشته باشند بردارهای پشتیبان (SVs)[11]، در شکل۱ نشان داده شده است، نامیده می‌شوند. لذا آموزش SVM تنها بر مبنای بردار های پشتیبان استوار است و از بقیه داده‌ها می‌توان صرف نظر نمود.



شکل۱: رده‌بندی جداکننده خطی(آلام[26]، ۲۰۰۰)

[9] Structural Risk Minimization
[10] Empirical Risk Minimization
[11] support vectors

بردارهای پشتیبان در بین دو خط موازی قرار می‌گیرند که موازی خط جدا کننده می‌باشند. معادله این دو خط به صورت معادلات (۳) و (۴) می‌باشد.

$$\mathbf{w}.\mathbf{x}+b=1 \qquad (\text{Class A}), \tag{۳}$$

$$\mathbf{w}.\mathbf{x}+b=-1 \qquad (\text{Class B}). \tag{۴}$$

هنگامی که SVM آموزش دید، تابع تصمیم در معادله (۵)[۱۲] مشخص می‌سازد که یک نمونه به کدام مرز تصمیم یا دسته تعلق دارد.

$$f(x)=\operatorname{sgn}(\mathbf{w}.\mathbf{x}+b) \tag{۵}$$

آموزش SVM بر مبنای بهینه‌سازی تابع معادله (۶) استوار است. که در آن l تعداد داده‌های آموزشی و α_i ضرایب لاگرانژ می‌باشند که با در نظر گرفتن قیود رابطه (۷)، حاصل گردیده است.

$$L=\frac{1}{2}\|\mathbf{w}\|^2-\sum_{i=1}^{l}\alpha_i y_i(\mathbf{w}.\mathbf{x}_i+b)+\sum_{i=1}^{l}\alpha_i, \tag{۶}$$

$$y_i(\mathbf{w}.\mathbf{x}_i+b)\geq 1, \tag{۷}$$

بنابراین حل مساله به صورت زیر ممکن می‌گردد. که v_i برابر با $\alpha_i y_i$ می‌باشد.

$$\mathbf{w}=\sum_{i=1}^{l}\alpha_i y_i \mathbf{x}_i=\sum_{i=1}^{l}v_i\mathbf{x}_i, \tag{۸}$$

باجایگذاری معادله(۸) درمعادله(۵) معادله (۹) حاصل می‌گردد.

$$f(x)=\operatorname{sgn}\left(\sum_{i=1}^{l}v_i(\mathbf{x}_i.\mathbf{x})+b\right) \tag{۹}$$

اگر مجموعه داده‌ها به صورت بهینه تفکیک شوند به این معنی است که ابر صفحه بدون هیچ خطایی داده‌ها را تقسیم نموده و فاصله بین نزدیک‌ترین بردار به ابر صفحه حداکثر می‌باشد (گان[۲۵]، ۱۹۹۸). در مورد داده‌های تفکیک ناپذیر خطی، باید ابر صفحه به نحوی تعریف گردد که اجازه جدایی پذیری خطی در ابعاد بالاتر را فراهم آورد(این امر متناظر با ابر صفحه غیر خطی می‌باشد). در این حالت می‌توان با نگاشت غیرخطی بردار ورودی به یک فضای ویژگی با ابعاد بالاتر که از دیدگاه ورودی و خروجی پنهان است، یک ابر صفحه بهینه برای تفکیک داده ها به دست می آید. به همین منظور از توابع هسته $K(x_i,x)$ استفاده می‌شود. با اصلاح معادله (۹)، معادله (۱۰) برای حالت داده‌های تفکیک ناپذیر خطی حاصل می‌گردد.

---

[۱۲] منظور از sgn، تابع sign می‌باشد.

آکادمی داده
dataacademy.ir

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{l} v_i K(\mathbf{x}_i, \mathbf{x}) + b\right) \quad (۱۰)$$

توابع هسته متعددی همچون تابع پایه شعاعی، چندجمله‌ای، سیگموئید و ... وجود دارد که روابط مربوطه به‌ترتیب در معادلات (۱۱، ۱۲، ۱۳) نشان داده شده است(ارپی‌برت و میکا[۲۷]، ۲۰۰۱). در این مقاله، یکی از رایج‌ترین آن‌ها توابع پایه‌ی شعاعی (RBF)[۱۳] بکارگرفته شده است.

$$K(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \quad (۱۱)$$

$$K(x, y) = (x.y + 1)^d \quad (۱۲)$$

(13)

$$K(x, y) = \tanh(v(x, y) + c)$$

۲-۲- الگوریتم ژنتیک در انتخاب ویژگی

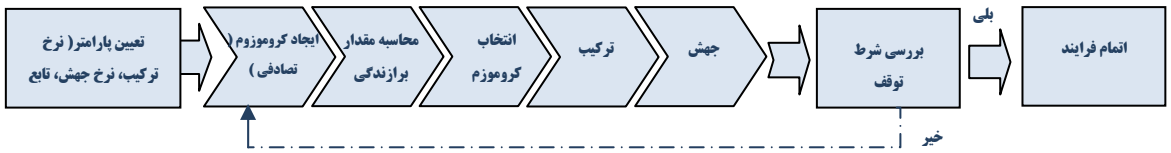
در طبقه‌بندی کلی، روش‌های انتخاب ویژگی به دو دسته فیلترها و پوشه‌ها[۱۴] تقسیم می‌شوند. فیلترها روش‌هایی از انتخاب ویژگی هستند که در آن‌ها راهکار انتخاب از مدل رده‌بندی‌کننده داده‌ها مستقل است و در طراحی آن از شاخص‌هایی همانند واریانس، همبستگی، آنتروپی و ... استفاده می‌شود (مانند PCA). دسته دوم، راهکارهای انتخاب ویژگی، پوشه‌ها هستند که در آن‌ها عمل انتخاب با یک طبقه‌بندی کننده یا الگوریتم هوشمند، جهت بیشینه/کمینه کردن یک معیار و شاخص همراه است. اگرچه مقایسه فیلترها و پوشه‌ها به منظور یافتن برتری یکی بر دیگری، پرسشی نادقیق و غیر علمی است، اما باور عموم بر آن است که چنانچه هزینه‌های زمانی/منابعی روش‌های پوشه را بتوان قبول کرد راهکارهای پوشه ارجح هستند(لین و همکاران[۲۸]، ۲۰۰۹). در این تحقیق نیز، جهت انتخاب ویژگی‌های مناسب مجموعه داده‌ها، از راهکار دوم (پوشه) استفاده می‌گردد.

الگوریتم ژنتیک، بر مبنای نظریه تکامل تدریجی داروین، توسط «جان هالند» (جان هالند[۲۹]، ۱۹۷۵) پایه‌گذاری شده است. روال بهینه‌یابی در الگوریتم ژنتیک بر اساس یک روند تصادفی هدایت شده استوار است. در این روش، ابتدا برای تعدادی از جواب‌ها که جمعیت[۱۵] نامیده می‌شود، مجموعه‌ای از پارامترهای هدف به صورت تصادفی تولید می‌شود. در واقع هر عضو از جمعیت (جواب) به صورت رشته ترکیبی از متغیرها (ویژگی‌ها نشان داده می‌شود که به هر یک از متغیرها، ژن و

---

[13] Radial Basis Functions
[14] wrappers
[15] population

به رشته مرکب از ژن‌ها جهت ایجاد یک جواب تصادفی، کروموزوم گفته می‌شود. پس از اجرای برنامه شبیه‌ساز عددی که معرف انحراف معیار و یا برازش آن مجموعه از اطلاعات است، (مقدار برازندگی)[۱۶]، به آن عضو از جمعیت مذکور نسبت داده می‌شود. این عمل برای تک‌تک اعضای ایجاد شده تکرار می‌شود، سپس با فراخوانی عملگرهای الگوریتم ژنتیک از جمله ترکیب[۱۷]، جهش[۱۸]، انتخاب نسل بعد شکل می‌گیرد و این روال تا برآورده شدن یکی از شروط توقف ادامه خواهد یافت. ساختار فوق، به اختصار در نمودار ذیل نشان داده شده است (بابگلو و فیندیک[۳۰]، ۲۰۱۰)



شکل۲: ساختار الگوریتم ژنتیک (بابگلو و فیندیک[۲۹]، ۲۰۱۰)

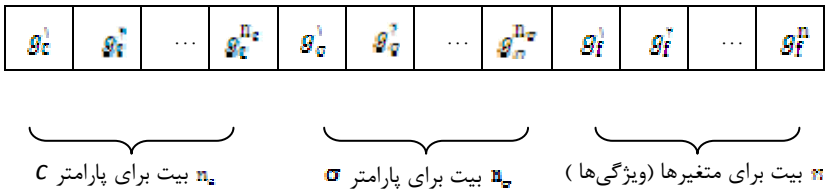
۳-چارچوب مدل: الگوریتم‌های هیبریدی GA-SVM

در این روش، ماشین بردار پشتیبان غیرخطی با توابع پایه شعاعی (RBF)[۱۹] به‌عنوان مدل رده‌بندی درنظر گرفته شده است. همانطور که در معادله (۱۱) نشان داده شده است، σ پهنای هسته‌ی RBF می‌باشد و پارامتر مهمی در عملکرد رده‌بندی محسوب می‌گردد. همچنین در داده‌های تفکیک‌ناپذیر، یک همپوشانی بین دسته‌ها وجود دارد، لذا محدوده پارامتر v_i باید به‌منظورکاهش تاثیر همپوشانی محدوده‌های تعریف شده‌ی بردارهای پشتیبان مقید گردد (یعنی $|v_i| \leq C$). در داده‌های تفکیک‌پذیر، مقدار C بینهایت می‌باشد در حالی که در داده‌های تفکیک‌ناپذیر این پارامتر ممکن است بر حسب خطاهای مجاز در آموزش تغییر کند. مقادیر بالای C اجازه خطای کمتر و مقادیر پایین C اجازه بروز خطای بیشتر در مساله را فراهم می‌آورد (لین و همکاران[۲۷]، ۲۰۰۹). بنابراین تخمین صحیح پارامترهای مذکور (C و σ) تاثیرقابل توجهی بر عملکرد مدل دارد. در مدل ترکیبی، انتخاب ویژگی و تعیین پارامترهای مدل به صورت همزمان و بدون کاهش دقت مدل رده‌بندی، توسط الگوریتم ژنتیک انجام می‌شود. الگوریتم ژنتیک استفاده شده در این رویکرد باینری است و مقادیر ژن‌ها در هر کروموزوم به صورت صفر یا یک است. با توجه به این که در هر جواب (کروموزوم) علاوه بر n متغیر به عنوان ویژگی‌های مدل، ارزش دو پارامتر مدل با مقادیر پیوسته را نیز می‌بایست تعیین نماید، با در نظر گرفتن دامنه تغییرات هر پارامتر، این مقادیر به مبنای ۲،

---

16 fitness value
17 crossover
18 mutation
19 Radial Basis Functions

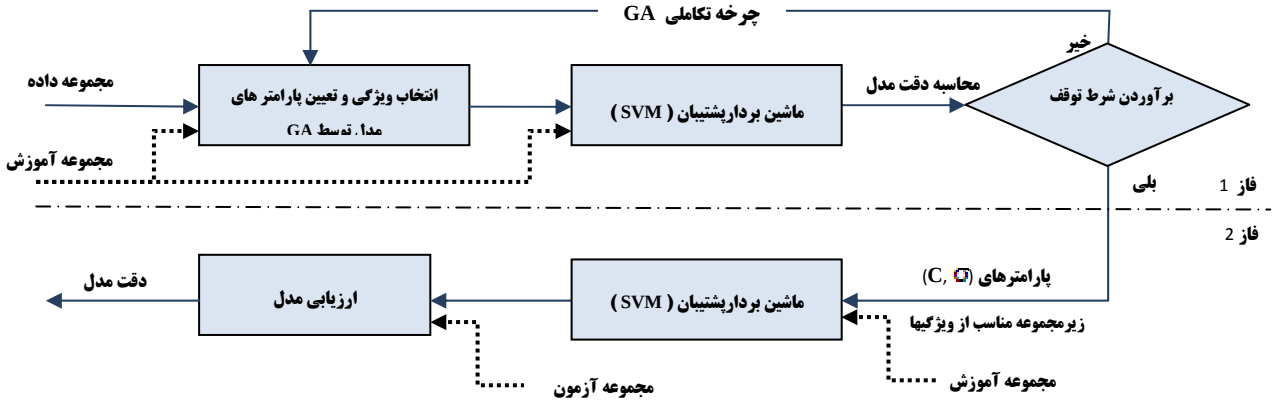
انتقال داده می‌شوند و به همان تعداد، بیت (ژن) در کروموزوم به آن‌ها تخصیص داده می‌شود. بنابراین همان طور که در شکل ۳ مشخص شده است n_c بیت (ژن) به پارامتر c، n_σ بیت به پارامتر σ و n بیت برای ویژگی‌های داده در نظر گرفته شده است.



شکل ۳: نمای کلی کروموزوم استفاده شده

ارزیابی تابع برازندگی[20] در الگوریتم ژنتیک بر اساس دقت مدل رده‌بندی مجموعه داده‌های آزمون صورت می‌گیرد. دقت مدل به کمک روش k-fold تعیین می‌گردد. بدین صورت که مجموعه داده آموزش به k بخش تقسیم شده، *k-1* بخش برای ساخت مدل و یک بخش باقی مانده جهت ارزیابی مجموعه آموزش به کارگرفته شده است. با k بار تکرار این رویه، هر بخش یک مرتبه در تست و *k -1* مرتبه در مدل‌سازی سهیم بوده و در انتها میانگین دقت در *k* مرحله به عنوان دقت مجموعه آموزش در نظر گرفته می‌شود. از آن جایی که آموزش الگوریتم ژنتیک، یک فرایند تکاملی تصادفی است، استفاده از این روش و در نظرگرفتن میانگین دقت k مرحله به‌عنوان دقت نهایی، میزان تغییرات تصادفی فرایند آموزش را در الگوریتم ژنتیک کاهش می‌دهد.

dataacademy.ir



شکل ۴: الگوریتم GA-SVM (هوانگ و همکاران[30]، ۲۰۰۶)

[20] Fitness

پس از برآورده‌شدن شرط توقف، ماکزیمم تکرار و یا دقت مدل رده‌بندی، فرایند الگوریتم ژنتیک پایان می‌یابد و مقادیر پارامترهای مدل جهت ادامه فرایند تعیین می‌گردند. در مرحله بعد به کمک SVM، با دراختیار داشتن ویژگیهای انتخاب شده و مقادیر (σ و C) و براساس مجموعه داده آموزش، مدل رده‌بندی ساخته و نتایج ارزیابی دقت مدل با توجه به مجموعه داده آزمون مشخص می‌گردد (هوانگ و همکاران[۳۱]، ۲۰۰۶) . (مطابق شکل ۴)

٤- پیش‌بینی ریزش مشتری با استفاده از الگوریتم GA-SVM

در این بخش با دراختیار مجموعه داده اطلاعات مشتریان فروشگاه Mondrian در طی دو سال، مدل رده‌بندی به منظور پیش‌بینی ریزش مشتریان ساخته و عملکرد آن مورد ارزیابی قرار می‌گیرد. تعداد مشتریان مورد بررسی ۱۹۰ مورد می‌باشد که به صورت تصادفی از میان مجموع کل مشتریان فروشگاه انتخاب گردیده شده است. از این تعداد، ۱۳۸ مشتری در سال دوم ریزش نموده و ۵۲ مشتری هم چنان همان فروشگاه را جهت خرید انتخاب نمودند. ویژگی‌های درنظر گرفته دربرگیرنده مشخصات فردی هر مشتری و اطلاعات مربوط به خرید وی در طول دو سال از فروشگاه می‌باشد. این ویژگی‌ها شامل سن، منطقه سکونت، جنسیت، وضعیت تاهل، میزان درآمد سالیانه، تعداد فرزند، تعداد فرزند در خانه، تحصیلات، نوع حساب بانکی، شغل، وضعیت تملک، تعداد خودرو، تعداد مراجعه به فروشگاه، تعداد محصولات خریداری شده، ارزش محصولات خریداری شده، میزان سودآوری حاصل از خرید، نوع محصولات خریداری شده و تنوع آنها می‌باشند. متغیر خروجی، به صورت باینری (ریزش مشتری و یا عدم ریزش مشتری) تعیین شده است. به منظور انتخاب تابع هسته مناسب، مدل رده‌بندی تحت توابع پایه‌ای شعاعی، چندجمله‌ای، سیگموئید و نیز SVM خطی بر روی مجموعه داده پیاده‌سازی شده است. مطابق با جدول (۱)، ماشین بردار پشتیبان غیرخطی با تابع پایه شعاعی، به دلیل برخورداری از دقت عملکرد بالاتر در پیش‌بینی ریزش مشتری به خصوص در مورد covering rate[۲۱]، جهت استفاده در الگوریتم ترکیبی GA-SVM انتخاب گردیده است.

جدول۱ : ارزیابی دقت مدل ماشین بردار پشتیبان تحت توابع مختلف

تابع کرنل	پارامتر ۱	پارامتر ۲	accuracy rate	hit rate	covering rate	lift coefficient
Linear SVM	-	-	۰.۶۶۳۲	۰.۷۲۲۹	۰.۸۶۹۵	۰.۹۱۳۰
SVM-RBF	$\sigma = 1$	C=1	۰.۷۲۶۳	۰.۷۳۱۲	۰.۹۸۵۵	۱
SVM-Polynomial	d=3	C=1	۰.۶۴۲۱	۰.۷۱۶۰	۰.۸۴۰۶	۰.۸۸۴۰
SVM-Sigmoid	[1, -1]	C=1	۰.۶۶۳۱	۰.۷۱۲۶	۰.۸۹۸۵	۰.۹۱۳۰

[۲۱] **در مسائل پیش‌بینی ریزش مشتری، خسارت ناشی از پیش‌بینی اشتباه یک مشتری ریزش دارای اهمیت بیشتری نسبت به پیش‌بینی اشتباه یک مشتری وفادار می‌باشد.**

معیارهای ارزیابی دقت مدل، با توجه به جدول ۲ و بر اساس روابط accuracy rate = (A+D)/(A+B+C+D)، hit rate = A/(A+C)، A/(A+B) = covering rate و lift coefficient = accuracy rate / customer churn rate محاسبه شده است. (کانگ و پی‌جی[1]، ۲۰۰۸)

جدول۲ : پارامترهای ارزیابی مدل(کانگ و پی‌جی[1])

		پیش‌بینی مدل	
		تعداد مشتریان ریزش	**تعداد مشتریان وفادار**
واقعیت	**تعداد مشتریان ریزش**	**A**	**B**
	تعداد مشتریان وفادار	**C**	**D**

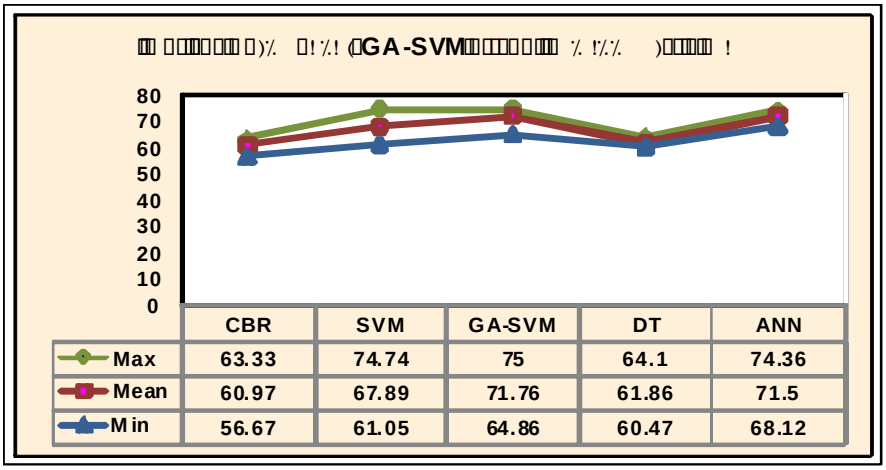
به منظور تخمین پارامترهای مدل (Cو σ) و انتخاب ویژگی توسط الگوریتم ژنتیک، هر کروموزوم به صورت یک رشته ۳۹ بیتی تعیین شده است. پارامتر C در محدوده [۳۵۰۰۰-۰.۰۱] و σ در محدوده [۳۲-۰.۰۰۰۱] در نظر گرفته شده است (لین و همکاران[27]، ۲۰۰۳). که این مقادیر به حالت باینری انتقال داده می‌شوند. ۳۰٪ مجموعه داده جهت آزمون و ۷۰٪ مابقی جهت ساخت مدل تقسیم‌بندی گردیده است. به منظور اجتناب از انطباق بیش از حد مدل ساخته شده با داده‌های مجموعه آموزش، از روش K-fold، استفاده شده است. مقدار بهینه k، با توجه به جدول (۳)، برابر با ۴ در نظر گرفته شده است.

جدول ۳: دقت مدل به ازاء مقادیر مختلف k

K	روش GA-SVM			
	پارامتر C	پارامتر σ	دقت مدل	تعداد ویژگی انتخاب شده
۳	۱۵۱۲	۰.۹۴	۶۷.۸۶	۱۰
۴	۱۵۳۲	۲.۴۴	۷۳.۲۱	۷
۵	۵۰۸	۱.۸۴	۶۴.۲۸	۹
۶	۳۶۸	۰.۹۱	۶۹.۶۴	۸
۷	۱۴۵۸	۱.۱۲	۶۷.۸۶	۹

به منظور ارزیابی عملکرد الگوریتم ترکیبی GA-SVM در پیش‌بینی ریزش مشتری، نتایج حاصله با دقت رده‌بندی مدلهای شبکه‌عصبی‌مصنوعی (ANN)، درخت تصمیم (DT)، استدلال مبتنی بر مورد (CBR) و ماشین بردار پشتیبان (SVM) مقایسه شده است. (مطابق با نمودار۱)

جهت بررسی دقیق‌تر وجود اختلاف مابین الگوریتم ترکیبی GA-SVM و هریک از مدلهای رده‌بندی بیان شده در نمودار(۱)، از آزمون t-test استفاده شده است. مقادیر آماره P-Value حاصل از مقایسه GA-SVM با هر یک از روشهای DT، SVM و CBR به ترتیب برابر با ۰، ۰.۰۴۸ و ۰ می‌باشد و مشخص گردید که در سطح اطمینان ۹۹٪، دارای اختلاف معناداری نسبت به روشهای CBR و DT و درسطح ۹۵٪ اختلاف معناداری نسبت به روش SVM دارد. همانطور که مشخص شده است، با توجه به اهمیت و حساسیت مرحله انتخاب ویژگی بردقت عملکرد مدل، تفاوت قابل توجهی نسبت به روش SVM و GA-SVM مشاهده می‌شود. همچنین قابل ذکر می‌باشد طبق آزمون انجام شده، اختلاف معناداری نسبت به نتایج دقت رده‌بندی دو مدل GA-SVM و ANN در پیش‌بینی ریزش مشتری مجموعه داده به کار گرفته شده وجود نداشته است.



۳- جمع‌بندی

اشباع بازار، تشدید رقابت مابین شرکت‌های ارائه دهنده خدمات و محصولات و هزینه‌های بالای جذب مشتری جدید در مقایسه با حفظ مشتریان قدیمی، از دلایل اصلی گرایش شرکت‌ها به حفظ مشتری می‌باشد. از آن جایی که یکی از استراتژی‌های اصلی حفظ مشتری، شناسایی مشتریان مستعد ترک می‌باشد، مدلهای تحلیلی و پیش‌بینی قدرتمندی در این زمینه به کار گرفته می‌شود. در این مقاله از ماشین بردار پشتیبان (SVM) به منظور پیش‌بینی ریزش مشتری بهره گرفته شده است. SVM، با نگاشت غیرخطی متغیرهای ورودی به فضای ویژگی با ابعاد بالاتر، مسائل پیچیده را به توابع جداپذیر ساده‌تر تفکیک می‌کند. و به این دلیل که مبتنی بر کمینه‌سازی ریسک ساختاری (SRM) می‌باشد، با حداقل نمودن حد بالای ریسک واقعی، عملکرد بسیار مناسبی در مواجهه با مجموعه داده‌های جدید دارد. اعتبارسنجی مدل بکارگرفته، با مقایسه آماری دقت مدلهای رده‌بندی ANN، DT و CBR صورت گرفته و نشان داده شد که دقت عملکرد بهتری نسبت به سایر

روشهای مورد بررسی دارد. هم چنین دارای اختلاف معناداری نسبت به دو روش DT و CBR می‌باشد. تخمین پارامترهای مدل و انتخاب ویژگی به صورت همزمان توسط الگوریتم ژنتیک باینری انجام گرفت. برتری قابل توجه دقت رده‌بندی مدل ترکیبی نسبت به ماشین بردار پشتیبان، نشان از حساسیت بالای مدل پیش‌بینی نسبت به متغیرهای ورودی دارد. در این مطالعه از راهکار دوم انتخاب ویژگی که در آن انتخاب با یک الگوریتم هو شمند صورت می‌گیرد، استفاده شده است. با این وجود نیاز به تحقیقات گسترده‌تر در رابطه با تحلیل حساسیت مدل نسبت به روشهای مختلف انتخاب ویژگی و بهینه‌سازی پارامترها و نیز به کارگیری سایر تکنیک‌های پیشرفته و دقیق داده‌کاوی در حوزه پیش‌بینی ریزش مشتری ضروری تلقی می‌گردد. .

۴- مراجع

Acir, N. 2006. A support vector machine classifier algorithm based on a perturbation method and its application to ECG beat recognition systems. Expert Systems with Applications, 31(1): 150–158.

Alam, P., Booth, D., Lee, K. and Thordarson, T.2000. The use of fuzzy clustering algorithm and self-organization neural networks for identifying potentially failing banks: An experimental study", Expert Systems with Applications, 18:.185–199.

Babaoglu.I , Findik.O.2010.A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine , Expert Systems with Applications, 37:3177–3183.

Bicego, M., Grosso, E. and Tistarelli, M. 2005. Face authentication using one-class support vector machines. Lecture Notes in Computer Science, 3781: 15–22.

Bin, L., Peiji, S. and Juan, L. 2007. Customer churn prediction based on the decision tree in personal handyphone system service. International conference on service systems and service management .

Bratko, A., and Filipic, B. 2006. Exploiting structural information for semi-structured document categorization. Information Processing and Management, 42(3):679–694.

Burges.C.J.1998.A tutorial on support vector machines for pattern recognition", Data mining and knowledge discovery, (2):121–167.

Chen, K.-Y. and Wang, C.-H. 2007. A hybrid SARIMA and support vector machines in forecasting the production values of the machinery industry in Taiwan. Expert Systems with Applications, 32(1): 254–264.

Chen, X. J., Harrison, R and Zhang, Y. Q. 2005. Multi-SVM fuzzy classification and fusion method and applications in bioinformatics. Journal of Computational and Theoretical Nanoscience, 2(4): 534–542.

Coussement.K, Vanden, Poel.2008. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques, Expert Systems with Applications 34 : 313–327.

E Xu, Shao.L, Gao.X, Zhai.B.2006. An algorithm for predicting customer churn via BP neural network based on rough set, IEEE Asia-Pacific Conference on Services Computing, (APSCC’06) : 47-50.

Glotsos, D., Tohka, J. and Ravazoula, P. 2005. Automated diagnosis of brain tumours astrocytomas using probabilistic neural network clustering and support vector machines. International Journal of Neural Systems, 15(1–2): 1–11.

Gunn.S.R. 1998.Support vector machines for classification and regression", ISIS technical report, (14).

Holland, J.1975. " Adaptation in natural and artificial systems, The Michigan University Press.

Huang.B, Kechadi.T, Buckley.B, Kiernan.G, Keogh.E, Rashid.T.2010.A new feature set with new window techniques for customer churn prediction in land-line telecommunications”, Expert Systems with Applications 37 : 3657–3665.

Huang.CH, Chen.M, Wang.CH .2006.Credit scoring with a data mining approach based on support vector machines", Expert Systems with Applications.

Jing.Z , Xing-hua.D .2008.Bank Customer Churn Prediction Based on Support Vector Machine: Taking a Commercial Bank’s VIP Customer Churn as the Example,IEEE.

Kang, C. Pei-ji, SH.2008, Customer Churn Prediction Based on SVM-RFE, International Seminar on Business and Information Management.
KhakAbi.S, Gholamian.M, Namvar.M.2010.Data Mining Applications in Customer Churn Management, International Conference on Intelligent Systems, Modelling and Simulation.
Kim.H Yoon.CH.2004.Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market", Telecommunications Policy 28 : 751–765.
Kim, S. K., Yang, S., & Seo, K. S. 2005. Home photo categorization based on photographic region templates. Lecture Notes in Computer Science, 3689: 328–338.
Li, S.-T., Shiue, W., & Huang, M.-H.2006. The evaluation of consumer loans using support vector machines. Expert Systems with Applications, 30(4): 772–782.
Lin.SH, Shiue.Y, Chen.SH, Cheng.H.2009.Applying enhanced data mining approaches in predicting bank performance: A case of Taiwanese commercial banks", Expert Systems with Applications , 11543–11551.
Lin.SH, Ying.K, Chen.SH, Lee.Z.2008. Particle swarm optimization for parameter determination and feature selection of support vector machines, Expert Systems with Applications, 1817–1824.
Lin, H.-T., Lin, C.-J.2003. A study on sigmoid kernels for SVM and the training of non- PSD kernels by SMO-type methods", Technical report, University of National Taiwan, 1-32 .
Ming.G, Hui-li.Z ,Yu-wei.L.2005.An analysis customer loss with bayesian networks method. Journal of Nanjing University of Posts and Telecommunications, (25):79-83.
Pendharkar.P.2009.Genetic algorithm based neural network approaches for predicting churn in cellular wireless network services, Expert Systems with Applications(36) : 6714–6720.
Rpbert K, Mika S.2001. An Introduction to Kernel Based Learning Algorithms [J], IEEE Transactions on Neural Networks, 12(2): 181-202.
Vapnik.V.1995. The Nature of Statistical Learning Theory, New York, NY: Springer, 1995.
Wang.N, Niu. D.2009.Credit Card Customer Churn Prediction Based on the RST and LS-SVM, IEEE.
Xin.Z, Yi.W, Hong-wang.CH.2009.A New Prediction Model of Customer Churn Based on PCA Analysis, The 1st International Conference on Information Science and Engineering (ICISE).
Zhao, Y., Li, B., & Li, X. 2005. Customer churn prediction using improved one-class support vector machine. Lecture Notes in Artificia Intelligence, 3584: 300–306.

dataacademy.ir

[1] Kang and Pei-ji

[2] Xin, Yi, and Hong wang

[3] Coussement and Vanden

[4] Huang et al.

[5] Bin, Peiji, and Juan

[6] E Xu et al.

[7] Ming, Hui-li, and Yu-wei

[8] Pendharkar

[9] KhakAbi, Gholamian, Namvar

[10] Chen ,Harrison, and Zhang

[11] Li, Shiue, and Huang

[12] Chen and Wang

[13] Acir

[14] Bratko and Filipic

[15] Kim, Yang, and Seo

[16] Bicego, Grosso, and Tistarelli

[17] Glotsos, Tohka, and Ravazoula

[18] Kim and Yoon

[19] Zha and Li

[20] Jing and Xing-hua

[21] Wang and Niu

[22] Lin et al.

[23] Vapnik

[24] Burges

[25] Gunn

[26] Alam

[27] Rpbert and Mika

[28] Lin et al.

[29] Holland

[30] Babaoglu and Findik

[31] Huang, Chen, Wang