مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره هفتم شماره دوم. ص ۱۳۲-۱۲۴، ۱۳۹۹

ارائه یک مدل برای پیش‌بینی هزینه طراحی نرم‌افزار سیستم اطلاعات بیمارستانی به کمک الگوریتم درخت تصمیم پیوسته

الهام توکل[1]*، ولی نوذری[2]، علی پیرزاد[3]، سید احسان امیرحسینی[4]، علی عبداللهی[5]

• دریافت مقاله: ۹۸/۵/۱۲ • پذیرش مقاله: ۹۸/۶/۷

مقدمه: منظور از تخمین هزینه نرم‌افزار سیستم اطلاعات بیمارستانی، برآورد هزینه و زمان مورد نیاز برای توسعه این نرم‌افزار سیستم اطلاعات بیمارستانی پیش از شروع پروژه است که تا پایان تولید و توسعه سیستم ادامه دارد. تخمین هزینه نرم‌افزار برای تولید سیستم اطلاعات بیمارستان، یکی از دغدغه‌های مهم مدیریت پروژه شرکت‌های حوزه سلامت، تلقی می‌شود. الگوهای تخمین هزینه که در مراحل اولیه ساخت پروژه با حداقل اطلاعات موجود از پروژه، هزینه ساخت سیستم را تخمین می‌زنند، سودمند و مورد نیاز هستند. روش تخمین هزینه مناسب، امکان کنترل مؤثر زمان و هزینه ساخت سیستم را فراهم می‌نماید.

روش: در این مطالعه گذشته‌نگر ۲۳ نرم‌افزار متن‌باز سیستم اطلاعات بیمارستانی انتخاب شد و میزان هزینه طراحی نرم‌افزار و ۱۶ متغیر مستقل از هر نرم‌افزار سیستم اطلاعات بیمارستانی استخراج شد. سپس داده‌ها به مجموعه آموزشی و تست تبدیل شدند و به کمک الگوریتم درخت تصمیم پیوسته یک مدل پیش‌بینی برای تخمین هزینه سیستم اطلاعات بیمارستانی طراحی گردید. سپس الگوریتم با چهار الگوریتم پیوسته دیگر مورد ارزیابی قرار گرفت.

نتایج: در این مطالعه با روش 10-Fold الگوریتم درخت تصمیم پیوسته اجرا گردید و جهت ارزیابی از دو پارامتر میانگین مربعات خطا و درصد میانگین خطای مطلق استفاده گردید و در مدل پیشنهادی به خطای ۳۱/۷۴ واحد در میانگین مربعات خطا و خطای ۱۷٪ برای درصد میانگین خطای مطلق به دست آمد.

نتیجه‌گیری: در این مطالعه نشان داده شد که مدل پیشنهادی دارای خطای قابل قبولی است که نسبت به روش‌های مشابه بهتر عمل کرده است و می‌توان از آن برای تخمین هزینه سیستم های اطلاعات بیمارستانی استفاده نمود.

کلید واژه‌ها: سیستم اطلاعات بیمارستانی، درخت تصمیم پیوسته، تخمین هزینه، داده‌کاوی

•ارجاع: توکل الهام، نوذری ولی، پیرزاد علی، امیرحسینی سید احسان، عبداللهی علی. ارائه یک مدل برای پیش‌بینی هزینه طراحی نرم‌افزار سیستم اطلاعات بیمارستانی به کمک الگوریتم درخت تصمیم پیوسته. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۹؛ ۷(۲): ۳۲-۱۲۴.

۱. آموزشکده فنی و حرفه‌ای سما، دانشگاه آزاد سلامی واحد بندرماهشهر، بندر ماهشهر، ایران
۲. گروه تربیت بدنی، واحد ارسنجان، دانشگاه آزاد اسلامی، ارسنجان، ایران
۳. گروه مدیریت دولتی، واحد یاسوج، دانشگاه آزاد اسلامی، یاسوج، ایران
۴. گروه مدیریت ورزش، واحد یاسوج، دانشگاه آزاد اسلامی، یاسوج، ایران
۵. آموزشکده فنی و حرفه‌ای سما، دانشگاه آزاد اسلامی، واحد بندر ماهشهر، بندر ماهشهر، ایران

*** نویسنده مسئول:** الهام توکل
آدرس: بندر ماهشهر، دانشگاه آزاد اسلامی، واحد بندر ماهشهر
• شماره تماس: ۰۶۱۵۲۳۷۶۳۷۰ • **Email:** elham.tavakol90@gmail.com

مقدمه

روزبه‌روز بر تعداد بیمارستان‌هایی که از سیستم‌های فناوری اطلاعات سلامت استفاده می‌کنند افزوده می‌شود در حال حاضر می‌توان گفت تمام بیمارستان‌ها از سیستم‌های اطلاعات سلامت استفاده می‌کنند. به‌کارگیری کامپیوتر در بیمارستان‌ها ابزار قدرتمندی برای درمان و مدیریت فراهم آورده و منجر به افزایش بهره‌وری در بیمارستان‌ها شده است. طی ۱۰ سال گذشته تمایل شدیدی به استفاده از سیستم‌های اطلاعات بیمارستانی موسوم به سیستم‌های (Hospital Information System) HIS به وجود آمده است [۱].

عموماً شرکت‌های تولیدکننده HIS دچار معضل افزایش هزینه‌ها بیش از میزان بودجه پیش‌بینی شده هستند. طبق تحقیقی که توسط مؤسسه Standish انجام شده به طور متوسط هزینه‌های واقعی یک پروژه نرم‌افزاری ۸۹٪ بیشتر از میزان بودجه تعیین خاتمه می‌یابند و تنها ۱۱٪ پروژه‌ها به موقع، طبق هزینه بودجه شده و با تمام قابلیت‌ها و ویژگی‌هایی که از اول مشخص شده بودند به اتمام می‌رسند. به دنبال این قضیه، نارضایتی مشتری، بی‌کیفیتی نرم‌افزارها و ناامیدی تولیدکنندگان به وجود می‌آید [۲].

یک امر ضروری در مراحل اولیه پروژه، برآورد زمان و فعالیت لازم برای تکمیل پروژه است. متأسفانه این موضوع یکی از مشکل‌ترین کارها در حوزه فناوری اطلاعات سلامت است. اغلب پروژه‌های نرم‌افزاری دچار مشکل کمبود مالی و زمان می‌شوند که یکی از دلایل آن برآوردهای اولیه اشتباه است [۳].

به فرآیند پیش‌بینی میزان فعالیت لازم برای ایجاد یک سیستم نرم‌افزار بیمارستانی برآورد بهای تمام شده HIS گفته می‌شود. بررسی داده‌های مربوط به پروژه‌های مختلف نشان داده است که روند بهای تمام شده با برخی پارامترهای قابل اندازه‌گیری همبستگی دارند. این مشاهدات به ارائه مدل‌های متعددی منتهی شده است که برای ارزیابی، پیش‌بینی و کنترل بهای تمام شده نرم‌افزار HIS می‌توانند مورد استفاده قرار بگیرند [۴].

به منظور جلوگیری از افزایش هزینه‌ها و زمان بیش از میزان بودجه پیش‌بینی شده یک پروژه نرم‌افزاری، مدل‌های برآورد بهای تمام شده متعددی به وجود آمده است؛ به دلیل وجود تحولات شدید در تولید نرم‌افزار ایجاد مدلی که برآوردهای دقیق از پروژه در اختیار استفاده کننده قرار دهد بسیار مشکل است، لذا یکی از مهم‌ترین اهداف صنعت نرم‌افزار ایجاد مدل‌های مفیدی است که منطبق بر چرخه عمر تولید نرم‌افزار (Software Development Life - Cycle) باشد و هزینه تولید یک محصول نرم‌افزاری را به دقت برآورد نماید [۵].

سه عامل در تعیین کل هزینه یک پروژه نقش دارد: بهای نرم‌افزار و سخت‌افزار به علاوه هزینه نگهداری، هزینه ایاب و ذهاب و آموزش و هزینه نیروی انسانی. برای اغلب پروژه‌ها عمده‌ترین هزینه، هزینه نیروی انسانی است. رایانه‌های پر قدرت مناسب برنامه‌نویسی نرم‌افزارها نسبتاً ارزان هستند اگر چه ممکن است به واسطه این‌که مراحل تولید نرم‌افزار در ایستگاه‌های کاری متعدد انجام می‌شود، هزینه ایاب و ذهاب زیادی نیاز باشد؛ اما این هزینه‌ها به نسبت هزینه نیروی انسانی بسیار ناچیز است علاوه بر این به کارگیری سیستم‌های مخابره الکترونیکی از قبیل ایمیل، وب‌سایت و ویدئو کنفرانس می‌تواند این هزینه را کاهش دهد.

در این پژوهش قرار است یک مدل پیش‌بینی هزینه طراحی HIS ارائه شود که بتواند هزینه طراحی این نرم‌افزار را تخمین بزند؛ در این پژوهش برای پیش‌بینی از الگوریتم درخت تصمیم پیوسته [۶] استفاده شد، علت انتخاب الگوریتم درخت تصمیم پیوسته قدرت آن برخورد با داده‌های پیوسته می‌باشد و مسئله مورد مطالعه نیز یک مسئله پیوسته است.

بررسی مطالعات و تحقیقات انجام شده در زمینه برآورد بهای تمام شده طی سال‌های ۱۹۸۹ تا ۲۰۱۴ نشان می‌دهد که رایج‌ترین موضوع این تحقیقات معرفی و ارزیابی مدل‌های برآورد بوده است؛ این بررسی نشان می‌دهد که اکثر مقالات به بررسی مدل‌های برآورد از نقطه نظر فنی پرداخته‌اند (جدول ۱) [۵]. همچنین این تحقیقات حاکی از آن است که مقالات ارائه شده غالباً به مطالعه رویکردهای برآورد مبتنی بر رگرسیون پرداخته‌اند. باید به این نکته توجه کرد که بیشتر مدل‌های پارامترهای رایج، مثل COCOMO، جزء این گروه قرار می‌گیرند. تقریباً نیمی از مقالات به ساخت، بهبود یا مقایسه مدل‌ها با مدل‌های مبتنی بر رگرسیون پرداخته‌اند [۲].

همچنین این تحقیقات حاکی از آن است که مقالات ارائه شده غالباً به مطالعه رویکردهای برآورد مبتنی بر رگرسیون پرداخته‌اند. تقریباً نیمی از مقالات به ساخت، بهبود یا مقایسه مدل‌های مبتنی بر رگرسیون پرداخته‌اند (جدول ۲) [۷]. طی مطالعاتی که انجام گرفته است مشخص گردید تا به حال مدلی جهت پیش‌بینی هزینه طراحی نرم‌افزار HIS تا به حال انجام نشده است.

جدول ۱: طبقه‌بندی موضوعی مطالعات برآورد بهای تمام شده نرم‌افزار[۵]

موضوع تحقیق	۱۹۹۹-	۲۰۰۰-۲۰۱۰	۲۰۱۰-۲۰۱۴	جمع
روش برآورد	۳۰(٪۵۱)	۹۶(٪۵۰)	۵۸(٪۵۰)	۱۸۴(٪۴۹)
تابع برآورد	۸(٪۱۴)	۷(٪۴)	۳(٪۳)	۱۸(٪۵)
بهینه‌سازی مدل	۳(٪۵)	۱۳(٪۷)	۴(٪۳)	۲۰(٪۵)
تعیین اندازه سیستم	۵(٪۸)	۳۹(٪۲۰)	۱۶(٪۱۴)	۶۰(٪۱۶)
مسائل سازمانی	۹(٪۱۵)	۱۵(٪۷)	۱۴(٪۱۲)	۴۸(٪۱۳)
ارزیابی عدم اطمینان مدل‌ها	۲(٪۳)	۱۰(٪۵)	۱۳(٪۱۱)	۲۵(٪۷)
ارزیابی عملکرد برآورد	۲(٪۳)	۸(٪۴)	۶(٪۵)	۱۶(٪۴)
ویژگی‌های مجموعه داده‌های مدل	۰	۱(٪۱)	۲(٪۲)	۳(٪۱)
سایر	۰	۳(٪۱/۵)	۱(٪۱)	۴(٪۱)

جدول ۲: سیر تحقیقات حول موضوع روش‌های برآورد بهای نرم‌افزار [۷]

موضوع تحقیق		۱۹۹۹-	۲۰۰۰-۲۰۱۰	۲۰۱۰-۲۰۱۴	جمع
	رگرسیون	۲۱(٪۳۳)	۷۶(٪۳۸)	۵۱(٪۳۲)	۱۴۸(٪۳۷)
	برآورد بر اساس مقایسه	۱(٪۲)	۱۵(٪۷)	۱۵(٪۹)	۳۱(٪۸)
	قضاوت تجربی	۷(٪۱۱)	۱۳(٪۶)	۲۱(٪۱۳)	۱۵(٪۴)
	تجزیه کار	۳(٪۵)	۵(٪۲)	۴(٪۳)	۱۲(٪۳)
توابع نقطه‌ای		۷(٪۱۱)	۴۷(٪۲۳)	۱۴(٪۹)	۶۸(٪۱۷)
	طبقه‌بندی و رگرسیون	۰(٪۰)	۵(٪۲)	۹(٪۶)	۱۴(٪۴)
	شبیه‌سازی	۲(٪۳)	۴(٪۲)	۴(٪۳)	۱۰(٪۳)
	شبکه‌های عصبی	۰(٪۰)	۱۱(٪۵)	۱۱(٪۷)	۲۲(٪۶)
	تئوری	۲۰(٪۳۲)	۱۴(٪۷)	۵(٪۳)	۳۹(٪۱۰)
	روش بیز	۰(٪۰)	۱(٪۱)	۶(٪۴)	۷(٪۲)
	ترکیب روش‌های مختلف	۰(٪۰)	۳(٪۱)	۲(٪۱)	۵(٪۱)
	سایر	۲(٪۳)	۷(٪۳)	۱۶(٪۱۰)	۲۵(٪۶)

روش

در این مطالعه از ۲۳ نرم‌افزار HIS متن‌باز (Source Open) که در این حوزه وجود دارد، انتخاب شد؛ لذا نمونه‌گیری انجام نشده است. از این ۲۳ نرم‌افزار ۱۶ متغیر مستقل و یک متغیر وابسته پیوسته که هزینه نهایی نرم‌افزار HIS است. استخراج گردید جزئیات متغیرهای استخراج شده در جدول ۳ مشاهده می‌شود. متغیرهای مستقل و بازه این مقادیر بر اساس داده‌ها استاندارد ناسا استخراج شده است [۸]. همچنین پارامترهای هر یک از این ۲۳ نرم‌افزار HIS بر مبنای استاندارد مهندسی نرم‌افزار Pressman محاسبه شده است [۹]. این داده‌ها می‌بایست مورد تفسیر قرار بگیرد و مشخص شود که آیا این داده‌ها قابلیت تعمیم‌پذیر دارد یا خیر. تفسیر این داده‌ها در بخش ارزیابی انجام می‌پذیرد.

جدول ۳: جزئیات داده‌های سیستم‌های اطلاعات سیستم برای تخمین نرم‌افزار HIS

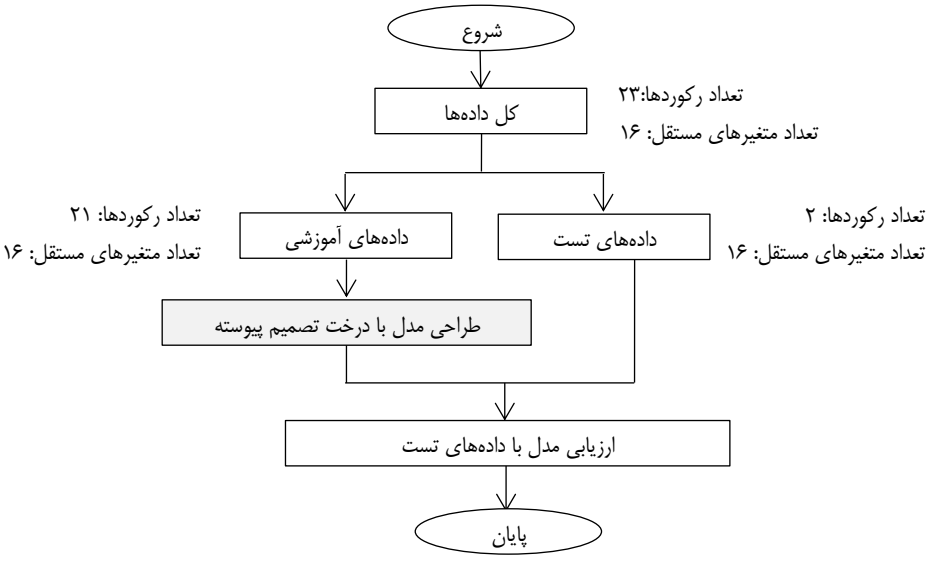
ردیف	نام متغیر	نقش متغیر	مقیاس اندازه‌گیری	روش اندازه‌گیری	بازه مقادیر
۱	قابلیت اطمینان نرم‌افزار	مستقل	رتبه‌ای	مشاهده‌ای	Low-Normal-High-Very High
۲	اندازه پایگاه داده	مستقل	رتبه‌ای	مشاهده‌ای	Low-Normal-High-Very High
۳	پیچیدگی فرآیندهای نرم‌افزار	مستقل	رتبه‌ای	مشاهده‌ای	Low-Normal-High-Very High-Extra High
۴	محدودیت زمانی پردازنده‌ها	مستقل	رتبه‌ای	مشاهده‌ای	Normal-High-Very High-Extra High
۵	محدودیت حافظه اصلی	مستقل	رتبه‌ای	مشاهده‌ای	Normal-High-Very High-Extra High
۶	نوسانات ماشین	مستقل	رتبه‌ای	مشاهده‌ای	Low-Normal-High
۷	زمان چرخش (زمان اجرا و بازگشت کامل به کاربر)	مستقل	رتبه‌ای	مشاهده‌ای	Low-Normal-High
۸	قابلیت آنالیز	مستقل	رتبه‌ای	مشاهده‌ای	Normal-High-Very High

۹	تجربه برنامه	مستقل	رتبه‌ای	مشاهده‌ای	Normal-High-Very High
۱۰	توانایی برنامه‌نویس	مستقل	رتبه‌ای	مشاهده‌ای	Normal-High-Very High
۱۱	ماشین مجازی	مستقل	رتبه‌ای	مشاهده‌ای	Low-Normal-High
۱۲	خبرگی زبان	مستقل	رتبه‌ای	مشاهده‌ای	Very Low-Low-Normal-High
۱۳	برنامه‌ریزی مدرن	مستقل	رتبه‌ای	مشاهده‌ای	Low-Normal-High-Very High
۱۴	استفاده از ابزارهای نرم‌افزاری	مستقل	رتبه‌ای	مشاهده‌ای	Low-Normal-High-Very High
۱۵	محدودیت برنامه	مستقل	رتبه‌ای	مشاهده‌ای	Low-Normal-High
۱۶	تعداد خطوط	مستقل	فاصله‌ای–پیوسته	مشاهده‌ای	[۲/۲ ،۴۲۳]
۱۷	هزینه نرم‌افزار	وابسته	فاصله‌ای–پیوسته	مشاهده‌ای	[۸/۴ ،۳۲۴۰]

الگوریتم پیشنهادی

در این قسمت الگوریتم پیشنهادی مطالعه که مبتنی بر درخت تصمیم پیوسته است ارائه گردید در مرحله اول همه داده‌ها با تمام ویژگی‌ها در نظر گرفته شد. سپس داده‌ها به قسمت آموزشی و تست تقسیم گردید. الگوریتم درخت پیوسته بر روی داده‌های آموزشی اجرا شد و یک مدل پیش‌بینی پیوسته مبتنی بر رگرسیون درخت تصمیم طراحی شد جهت ارزیابی مدل پیشنهادی از داده‌های تست استفاده شد نحوه ارزیابی در قسمت ارزیابی بیان گردید. فلوچارت طراحی مدل درخت تصمیم پیوسته در شکل ۱ مشاهده می‌شود. مدل پیشنهادی توسط نرم‌افزار متلب ۲۰۱۷ پیاده‌سازی گردیده است.

نحوه تقسیم داده‌های آموزشی و تست بر مبنای روش K-Fold است [۱۰]؛ در این نوع تقسیم‌بندی داده‌ها به K زیرمجموعه افراز می‌شوند. از این K زیرمجموعه، هر بار یکی برای تست و K-1 تای دیگر برای آموزش به کار رفتند. این روال K بار تکرار می‌شود و همه داده‌ها دقیقاً یک بار برای آموزش و یک بار برای تست به کار برده شدند در نهایت میانگین نتیجه K بار ارزیابی به عنوان یک تخمین نهایی برگزیده شد.



شکل ۱: مدل درخت تصمیم پیوسته پیشنهادی

همان طور که در شکل ۱ مشاهده شد، ۲۳ رکورد تحقیق به دو مجموعه آموزشی و تست تقسیم شد، ۲۱ رکورد برای داده‌های آموزشی و ۲ رکورد برای تست در نظر گرفته شد در شکل ۱ میزان تقسیم داده بر مبنای مقدار K برابر ۱۰ در نظر گرفته شده است.

در مرحله ارزیابی مدل پیشنهادی مورد ارزیابی قرار گرفت در مرحله اول داده‌های تحقیق مورد تفسیر قرار گرفت، سپس پارامترهای ارزیابی بیان شد و در ادامه الگوریتم پیشنهادی مورد ارزیابی قرار گرفت و با چهار الگوریتم دیگر ارزیابی شد (جدول ۴).

جدول ٤: محاسبه پارامترهای توصیفی برای داده مطالعه تخمین هزینه نرم‌افزار HIS

نام متغیر	میانگین	انحراف معیار	میانه	میانگین پیراسته	نما	حداقل	حداکثر	بازه	چولگی	کشیدگی	خطای استاندارد
قابلیت اطمینان نرم‌افزار	۱/۴۸	۰/۶	۱	۱/۴۶	۰	۰	۳	۳	۰/۳	−۰/۵	۰/۰۸
اندازه پایگاه داده	۰/۹۲	۱/۰۵	۱	۰/۷۷	۱	۰	۳	۳	۰/۷۷	−۰/۷۲	۰/۱۴
پیچیدگی فرایندهای نرم‌افزار	۱/۹۲	۰/۵۶	۲	۱/۹۸	۰	۰	۴	۴	−۰/۶	۵/۷۶	۰/۰۷
محدودیت زمانی پردازنده‌ها	۱/۵۷	۰/۸۷	۱	۱/۴۴	۰	۱	۴	۳	۱/۰۹	−۰/۳۹	۰/۱۱
محدودیت حافظه اصلی	۱/۵۳	۰/۹۱	۱	۱/۳۵	۰	۱	۴	۳	۱/۴۳	۰/۶۸	۰/۱۲
نوسانات ماشین	۰/۳	۰/۵۳	۰	۰/۲۱	۰	۰	۲	۲	۱/۵	۱/۳	۰/۰۷
زمان چرخش	۱/۰۷	۱/۲۹	۰/۵	۰/۹۶	۰/۵	۰	۳	۳	۰/۶۸	−۱/۳۲	۰/۱۷
قابلیت آنالیز	۱/۶۵	۰/۷۱	۲	۱/۵۶	۱	۱	۳	۲	۰/۵۹	−۰/۸۸	۰/۰۹
تجربه برنامه	۱/۷۳	۰/۷۱	۲	۱/۶۷	۱	۱	۳	۲	۰/۴۲	−۱	۰/۰۹
توانایی برنامه‌نویس	۱/۵	۰/۷	۱	۱/۳۸	۰	۱	۳	۲	۱/۰۲	−۰/۳۳	۰/۰۹
ماشین مجازی	۰/۸۳	۰/۴۲	۱	۰/۹	۰	۰	۲	۲	−۱/۰۴	۰/۹۱	۰/۰۵
خبرگی زبان	۱/۶۲	۰/۶۹	۲	۱/۷۵	۰	−۱	۲	۳	−۲/۰۸	۴/۵۷	۰/۰۹
برنامه‌ریزی مدرن	۱,۵	۰/۸۹	۲	۱/۵	۱	۰	۳	۳	−۰/۱۴	−۰/۸	۰/۱۲
استفاده از ابزارهای نرم‌افزاری	۱/۴۲	۰/۸۹	۱	۱/۳۵	۰	۰	۳	۳	۰/۸۲	−۰/۵۱	۰/۱۱
محدودیت برنامه	۰/۶۸	۰/۶۵	۱	۰/۶	۱	۰	۲	۲	۰/۴	−۰/۷۹	۰/۰۸
تعداد خطوط	۷۴/۵۹	۹۷/۱۷	۳۰/۵	۵۳/۲۴	۲۲/۴	۲/۲	۴۲۳	۴۲۰/۸	۱/۸۳	۲/۷۳	۱۲/۵۴
هزینه نرم‌افزار	۴۰۶/۴۱	۶۵۶/۹۷	۱۱۸/۸	۲۴۱/۴۸	۹۲/۴	۸/۴	۳۲۴۰	۳۲۳۱/۶	۲/۵۵	۶/۳۱	۸۴/۸۱

از اختلاف میانگین و میانگین پیراسته می‌توان به وجود یا عدم وجود داده‌ها خارج از بازه پی‌برد دو متغیر تعداد خطوط و هزینه نرم‌افزار دارای اختلاف بین مقدار میانگین و میانگین پیراسته هستند بدین معنی است که این دو ویژگی دارای مقادیر خارج از بازه هستند، ولی در ۱۵ ویژگی دیگر اختلاف زیادی بین میانگین و میانگین پیراسته وجود ندارد؛ لذا دارای مقادیر خارج از بازه نیستند.

با توجه به میزان چولگی و کشیدگی متغیرها جدول ۴ داده‌ها دارای توزیع نسبتاً نرمالی هستند همچنین با توجه به مقدار خطای استاندارد می‌توان تفسیر نمود که این داده‌ها قابلیت تعمیم به کل جامعه را دارند یا خیر؟ همان‌طور که در جدول ۴ مشاهده شد دو متغیر تعداد خطوط و هزینه نرم‌افزار دارای خطای استاندارد زیادی هستند و قابلیت تعمیم به کل جامعه را ندارند؛ ولی مابقی متغیرها دارای خطای استاندارد کمی هستند و قابلیت تعمیم به کل جامعه دارند.

پارامترهای ارزیابی

یکی از معیارهای ارزیابی که در اکثر مقالات از آن استفاده می‌شود میانگین مربعات خطا (Mean Squared Error)MSE است [۱۱]، میزان میانگین مربعات خطا از (رابطه ۱) محاسبه می‌گردد، $Value_{\text{Predict}_i}$ مقدار پیش‌بینی شده رکورد iام است، $Value_{\text{Actual}_i}$ مقدار واقعی رکورد i ام است و n تعداد کل رکوردها هست [۱۲].

$$MSE = \frac{1}{2}\sum_{i=1}^{n}(Value_{Actual_i} - Value_{\text{Predct}_i})^2 \quad \text{(رابطه ۱)}$$

مشکلی که پارامتر MSE دارد نمی‌توان آن را به صورت درصد بیان نمود همچنین اگر داده‌ها نرمال نشود MSE خطای تقریباً بالایی دارد؛ لذا می‌بایست داده‌ها نرمال شود. سپس میزان خطا محاسبه می‌گردد؛ لذا پارامتر دوم جهت ارزیابی بیان می‌شود این پارامتر بیان درصدی از پارامتر میانگین را دارد و درصد خطای مطلق است این پارامتر آماری (Mean MAPE(Absolute Percentage Error نام دارد رابطه این پارامتر در (رابطه ۲) بیان شده است [۱۲].

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{Value_{Actual_i} - Value_{\Pr edict_i}}{Value_{Actual_i}}\right| \quad \text{(رابطه ۲)}$$

جهت ارزیابی الگوریتم پیشنهادی از (رابطه ۲) استفاده شد و جهت مقایسه الگوریتم پیشنهادی به روش‌های دیگر از (رابطه ۱) استفاده شد؛ زیرا استاندارد مقالات برای مقایسه بین روش‌های مختلف (رابطه ۱) است هر چقدر پارامتر MAPE کمتر باشد، نشان دهنده مناسب بودن داده پیش‌بینی شده است.

ارزیابی درخت تصمیم پیوسته

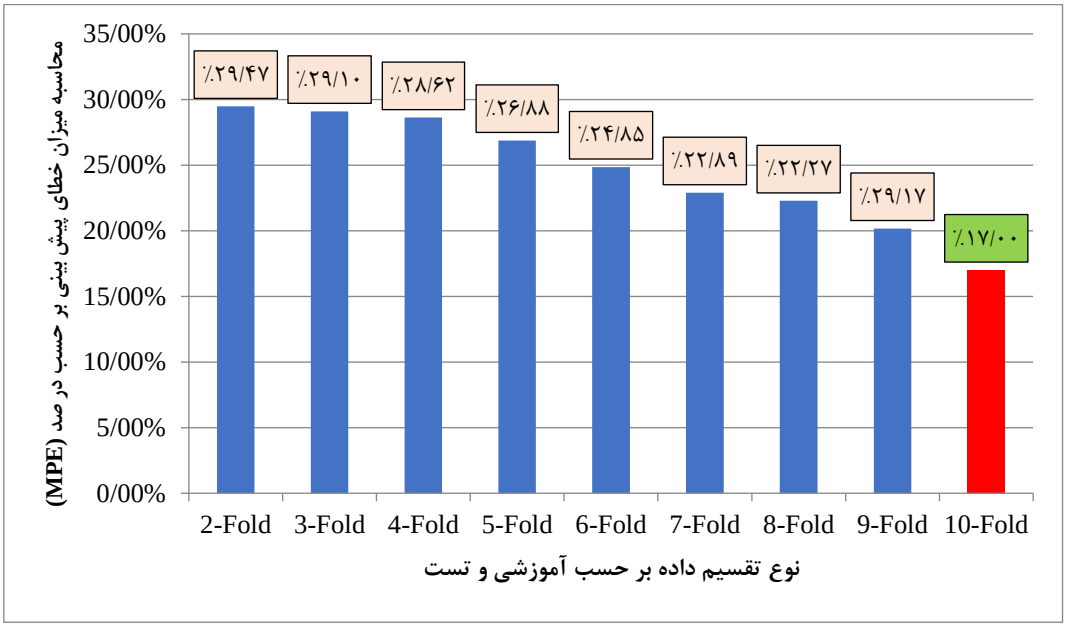
ارزیابی الگوریتم درخت تصمیم به این صورت است که داده‌ها به دو مجموعه آموزشی و تست تقسیم می‌شود [۱۳]. در این تحقیق از روش K-Fold استفاده شد برای تعیین مقدار بهینه K در این تحقیق مقدار K بین ۲ تا ۱۰ تغییر داده و مشخص می‌شود کدام مقدار K بهتر است.

در جدول ۵ و شکل ۲ خروجی الگوریتم درخت پیوسته با داده‌های تخمین پروژه مشاهده می‌شود. پارامتر ارزیابی MAPE است، الگوریتم ۱۰۰ بار اجرا می‌شود و میانگین ۱۰۰ بار گزارش شده است.

همین‌طور که در شکل ۲ و جدول ۵ مشاهده می‌شود مدل درخت تصمیم پیوسته دارای خروجی حداکثر درصد خطای ٪۲۹/۴۷۳۱ برای روش 2-Fold و دارای خروجی حداقل درصد خطای ٪۱۷/۰۰۲۳ برای روش 10-Fold است؛ لذا مقدار بهینه K برابر ۱۰ در نظر گرفته شد.

جدول ۵: خروجی الگوریتم درخت تصمیم پیوسته برای تخمین هزینه نرم‌افزارHIS

ردیف	میزان K	MAPE
۱	۲	٪۲۹/۴۷۳۱
۲	۳	٪۲۹/۰۹۷۷
۳	۴	٪۲۸/۶۱۷۴
۴	۵	٪۲۶/۸۸۱۷
۵	۶	٪۲۴/۸۵۴۸
۶	۷	٪۲۲/۸۹۴۹
۷	۸	٪۲۲/۲۷۳۷
۸	۹	٪۲۰/۱۶۸۹
۹	۱۰	٪۱۷/۰۲۲۳



شکل ۲: نمودار خروجی الگوریتم درخت تصمیم پیوسته برای تخمین هزینه نرم‌افزارHIS

با توجه به این که داده‌های مسئله پیوسته است از چهار الگوریتم رگرسیون بردار پشتیبان (Support Vector Regression)، رگرسیون چندگانه، نزدیک‌ترین همسایگی (Nearest Neighborhood) و شبکه عصبی مصنوعی (Artificial Neural Network) به منظور اطلاع از عملکرد این الگوریتم‌ها و مقایسه با روش پیشنهادی استفاده شد در تمامی الگوریتم‌های مورد مقایسه از روش K-Fold استفاده شده است و مقدار K برابر ۱۰ می‌باشد؛ در جدول ۶ نتایج ارزیابی الگوریتم‌های رگرسیون ماشین‌بردار، رگرسیون چندگانه، نزدیک‌ترین همسایگی و شبکه عصبی مصنوعی نمایش داده شده است. پارامتر ارزیابی در این قسمت میانگین مربعات خطا انتخاب شده است.

در جدول ۶ مقایسه الگوریتم درخت تصمیم با سه روش دیگر در این حوزه نشان داده شده است.

جدول ٦: مقایسه الگوریتم درخت تصمیم پیوسته با کارهای گذشته برای تخمین هزینه نرم‌افزار

نام الگوریتم	رگرسیون بردار پشتیبان	رگرسیون چندگانه	نزدیک‌ترین همسایگی	شبکه عصبی مصنوعی	الگوریتم پیشنهادی (درخت تصمیم پیوسته)
خروجی MSE	۴۵/۳۴	۴۹/۷	۳۴/۸۶	۳۲/۶۴	۳۱/۷۴

همان‌طور که در جدول ۶ مشاهده می‌شود الگوریتم درخت تصمیم پیوسته دارای خروجی بهتری نسبت به چهار الگوریتم دیگر دارد که نشان از مناسب بودن الگوریتم پیشنهادی است.

بحث و نتیجه‌گیری

مفهوم تخمین هزینه نرم‌افزار، هم‌زمان با شروع صنعت کامپیوتر از دهه ۴۰ میلادی مورد توجه قرار گرفته و همچنان پژوهش در این حوزه ادامه دارد. با این که تلاش، تنها در برگیرنده بخشی از هزینه‌های توسعه یک پروژه نرم‌افزاری در حوزه سلامت است؛ اما عامل اساسی برای تعیین هزینه محسوب می‌گردد؛ لذا در پژوهش‌های این حوزه، دو اصطلاح تخمین تلاش و تخمین هزینه به‌صورت معادل استفاده می‌شود. مدل تخمین هزینه نرم‌افزار HIS مناسب است که قبل از عقد قرارداد، دقت و اطمینان بالایی برای پیش‌بینی هزینه پروژه‌های این نرم‌افزار فراهم نماید. به علت ذات غیرقطعی تخمین و در جهت افزایش دقت، در این مطالعه از الگوریتم درخت تصمیم پیوسته استفاده شد.

هدف از این پژوهش، ارائه روشی برای دسته‌بندی مجموعه داده‌های تخمین نرم‌افزار سیستم اطلاعات سلامت با استفاده از الگوریتم درخت تصمیم پیوسته است؛ به‌طوری‌که بتوان از

طریق این الگوریتم، هزینه تمام شده نرم‌افزار HIS را با کارایی بالا پیش‌بینی نمود.

همان‌طور که در این پژوهش بیان گردید در داده‌های این مطالعه ضریب همبستگی بالایی وجود ندارد؛ لذا داده‌های این پژوهش دارای ارتباط غیرخطی بین متغیرهای مستقل و وابسته هستند؛ لذا از الگوریتم غیرخطی، درخت تصمیم پیوسته استفاده گردید. همچنین مشخص گردید که این داده‌ها با توجه به این که متغیر وابسته دارای خطای استاندارد بالایی است نتایج این پژوهش قابلت تعمیم ندارند، ولی می‌توان از خود مدل برای تخمین هزینه‌های داده‌های دیگر استفاده نمود.

در این مدل الگوریتم پیشنهادی با مقادیر 2-Fold تا 10-Fold مورد ارزیابی قرار گرفت و مشخص گردید که بهترین مقدار برای تقسیم داده‌ها به دو مجموعه آموزشی و تست مقدار 10-Fold است که خطای دسته‌بندی را به از ۲۹ درصد به ۱۷ درصد کاهش داده است. الگوریتم جهت همگرا شدن خروجی مسئله الگوریتم در هر مرحله ۱۰۰ بار اجرا گردید تا بتواند خروجی الگوریتم همگرا شود.

الگوریتم پیشنهادی نسبت به الگوریتم شبکه عصبی مصنوعی ۰/۹ واحد بهتر عمل کرده است و نسبت به الگوریتم نزدیک‌ترین همسایگی ۳/۱۲ واحد بهتر عمل کرده است و نسبت به الگوریتم رگرسیون بردار پشتیبان ۱۳/۶ واحد بهتر بود است همچنین نسبت به الگوریتم رگرسیون چندگانه ۱۷/۹۶ واحد بهتر عمل کرده است که نشان می‌دهد الگوریتم درخت تصمیم نسبت به روش‌های دیگر بهتر عمل کرده است.

از محدودیت‌های پژوهش به استفاده از نرم‌افزار متن‌باز برای داده‌های مسئله می‌توان نام برد به همین دلیل است این داده‌ها قابلیت تعمیم ندارند. اگر شرکت‌های طراح HIS اطلاعات ساختار برنامه خود را در اختیار قرار بدهند می‌توان این الگوریتم را برای آن اجرا نمود؛ ولی عموماً این شرکت‌ها به خاطر رقابت تجاری حاضر به همکاری نمی‌شود.

موضوعاتی زیر جهت انجام کارهای آینده پیشنهاد می‌گردد:

- از الگوریتم‌های فرا ابتکاری جهت افزایش درخت الگوریتم درخت تصمیم پیوسته
- استفاده از الگوریتم‌های کاهش ابعاد جهت افزایش سرعت اجرای الگوریتم
- استفاده از نرم‌افزار HIS غیر متن باز

تعارض منافع

در این مطالعه هیچ‌گونه تضاد منافعی وجود نداشت

References

1. Carvalho JV, Rocha Á, van de Wetering R, Abreu A. A Maturity model for hospital information systems. Journal of Business Research 2019;94:388-99. doi:10.1016/j.jbusres.2017.12.012

2. Boehm BW, Abts C, Brown AW, Chulani S, Clark BK, Horowitz E, Madachy R, Reifer DJ, Steece B. Software Cost Estimation with COCOMO II. 1st ed. USA: Prentice Hall; 2000.

3. Venkataiah V, Mohanty R, Pahariya JS, Nagaratna M. Application of ant colony optimization techniques to predict software cost estimation. In Computer Communication, Networking and Internet Security Singapore: Springer; 2017. p. 315-25. doi:10.1007/978-981-10-3226-4_32

4. Kaur R, Sharma ES. Various techniques to detect and predict faults in software system: survey. International Journal on Future Revolution in Computer Science & Communication Engineering 2018;4(2):330-6.

5. Agarwal BB, Tayal SP, Gupta M. Software Engineering And Testing: An Introduction (Computer Science). 1st ed. Canada: Jones & Bartlett Learning; 2009.

6. Krzanowski WJ, Hand DJ. ROC Curves for Continuous Data. 1st ed. USA: Chapman and Hall/CRC; 2009.

7. Saeed A, Butt WH, Kazmi F, Arif M. Survey of software development effort estimation techniques. In Proceedings of the 2018 7th International Conference on Software and Computer Applications; 2018 Feb 8; NewYork: Association for Computing Machinery; 2018. p. 82-6. https://doi.org/10.1145/3185089.3185140 8.

9. Pressman RS. Software Engineering: A Practitioner's Approach. 8th ed. Boston: McGraw-Hill Education; 2014.

10. Geng L, Hamilton HJ. Interestingness measures for data mining: A survey. ACM Computing Surveys (CSUR). 2006 Sep 30;38(3):9. https://doi.org/10.1145/1132960.1132963

11. Johnson RA, Bhattacharyya GK. Statistics: Principles and Methods. 8th ed. United States: John Wiley & Sons; 2019.

12. Rao CR, Miller JP, Rao DC. Epidemiology and Medical Statistics. Elsevier Science; 2007.

13. Berkhin P. A Survey of Clustering Data Mining Techniques. In Grouping Multidimensional Data. Singapore: Springer; 2006. p. 25-71.

Original Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2020; 7(2): 124-132

Providing a Model for Cost Estimation of Hospital Information System Software Design Using Continuous Decision Tree Algorithm

Tavakol Elham[1]*, Nowzari Vali[2], Pirzad Ali[3], Amirhosseini Seyed Ehsan[4], Abdolahi Ali[5]

• **Received:** 03 Aug 2019 • **Accepted:** 29 Aug 2019

Introduction: The cost estimation of a hospital information system software refers to estimating the cost and time required to develop the hospital information system software prior to the start of the project, which will continue until the end of production and development of the system. Estimating the cost of software to produce hospital information system is one of the major concerns of project management in health companies. Cost estimation models that estimate the cost of system construction in the early stages of project construction, with minimal information available from the project, are useful and needed. Selection of an appropriate cost estimation method enables efficient control of time and cost of system construction.

Method: In this retrospective study, 23 open source software projects for hospital information system were selected and the cost of software design and 16 independent variables of each hospital information system software were extracted. The data were then transformed into a test and training set and using a continuous decision tree algorithm, a prediction model was proposed to estimate the cost of designing a hospital information system. The algorithm was then evaluated with four other continuous algorithms.

Results: In this study, the continuous decision tree algorithm was implemented using the 10-fold method and two parameters including mean squared error and mean absolute percentage error were used for evaluation. In the proposed model, error of 74.31 units was obtained for the mean squared error and 17% for the mean absolute percentage error.

Conclusion: It was shown in this study that the proposed model had an acceptable error rate indicating that it performed better than similar methods and can be used to estimate the cost of hospital information systems.

Keywords: Hospital Information System, Continuous Decision Tree, Cost Estimation, Data Mining

• **Citation:** Tavakol E, Nowzari V, Pirzad A, Amirhosseini SE, Abdolahi A. Providing a Model for Cost Estimation of Hospital Information System Software Design Using Continuous Decision Tree Algorithm. Journal of Health and Biomedical Informatics 2020; 7(2): 124-32. [In Persian]

1. Sama Technical and Vocational Training College, Islamic Azad University, Bandar Mahshahr Branch, Bandar Mahshahr, Iran
2. Department of Physical Education, Arsanjan Branch, Islamic Azad University, Arsanjan, Iran
3. Department of Government Management, Yasooj Branch, Islamic Azad University, Yasooj, Iran
4. Department of Sport Management, Yasooj Branch, Islamic Azad University, Yasooj, Iran
5. Sama Technical and Vocational Training College, Islamic Azad University, Bandar Mahshahr Branch, Bandar Mahshahr, Iran

***Corresponding Author:** Elham Tavakol
Address: Islamic Azad University, Mahshahr Branch, Mahshar, Iran
• **Tel:** 06152376370 • **Email:** elham.tavakol90@gmail.com