Predict the Amount of Effort Required Doing the Computerized Project Using Decision Trees

پیش‌بینی میزان تلاش لازم برای انجام پروژهای کامپیوتری با استفاده از درخت‌های تصمیم‌گیری

Mehran Khosravi, Leila Rikhtechi, Shahram Andalibi

Abstract — One of the major aspects of computer project management is accurate information about the amount of time, effort and cost required to accomplish project. This problem is possible if to have information about the project and its development team. In order to model such as CoComo have been introduced, that's every project with a number of factors have been showed and using these factors to estimate the amount of effort required for the project. In this paper, by this factors and using data mining techniques with the help of decision trees a method for predicting the effort of software projects has been presented, that it is more accurate than similar method. The accuracy of this assessment criterion has been studied[1].

Keywords — effort predicting, data mining, decision trees, algorithmic and non-algorithmic methods.

۱. چکیده

همواره يكي از جنبه‌اي اصلي مديريت پروژه‌هاي كامپيوتري داشتن اطلاعات دقيق از ميزان زمان، تلاش و هزينه لازم براي انجام پروژه است. اين مسئله در صورتي امكان پذير است كه در مورد پروژه و تيم توسعه دهنده آن اطلاعاتي داشته باشيم. به اين منظور مدل‌هايي مانند مدل كوكومو معرفي شده‌اند كه هر پروژه را با تعدادي از فاكتورها نمايش مي‌دهند و با استفاده از اين فاكتورها ميزان تلاش لازم براي پروژه را تخمين مي‌كنند. در اين مقاله از اين فاكتور ها و با استفاده از تكنيك‌هاي داده كاوي و كمک گرفتن از درخت‌هاي تصميم گيري روشي براي پيش بيني ميزان تلاش در پروژه‌هاي نرم افزاري مطرح شده است كه نسبت به روش‌هاي مشابه از دقت بيشتري برخوردار است كه صحت اين مطلب از طريق معيارهاي ارزيابي مورد بررسي قرار گرفته است.

كلمات كليدي:
پيش بيني تلاش، داده كاوی، درخت‌های تصميم‌گيری، روش‌های الگوريتمی و غير الگوريتمی

۲. مقدمه

همواره در پروژه‌های بزرگ زمانبندی انجام پروژه یکی از مشکلات اصلی مدیر پروژه و تیم توسعه پروژه است. اصلی ترین معیار زمان بندی پروژه تعیین میزان تلاش لازم برای تکمیل پروژه است. میزان تلاش معمولا بر حسب واحد نفرماه تعیین می‌شود. رابطه‌ی مستقیمی بین میزان تلاش لازم برای تکمیل پروژه و زمان لازم برای تکمیل آن وجود دارد. و از روی زمان هم می‌توان هزینه را تخمین نمود. پس اصلی‌ترین مرحله در زمان بندی پروژه‌ها تعیین میزان تلاش لازم برای تکمیل پروژه است.

برای پیش بینی تلاش لازم برای انجام پروژه‌های نرم افزاری دو دسته روش اساسی وجود دارد: روش‌های الگوریتمی، روش‌های غیر الگوریتمی.

روش‌های غیر الگوریتمی هیچ گونه الگوریتم مشخصی ندارند و نتایج حاصل شده وابسته به مهارت افراد تخمین زننده و شرایط کلی پروژه بوده و نتایج حاصل شده قابل تکرار نمی‌باشد. اما در عین حال نسبت به روش‌های الگوریتمی تخمین تلاش با سرعت بیشتری حاصل می‌شود و درصورتیکه تیم تخمین کننده مهارت و تجربه کافی را داشته باشد تنایج مطلوبی حاصل می‌شود. نمونه این روش‌ها، روش مقایسه هزینه، کارشناس خبره، پارکینسون، هزینه تا موفقیت، پایین به بالا و بالا به پایین می‌باشد.

روش‌های الگوریتمی دارای الگوریتم‌ها و فرمول‌های کاملا مشخص و قابل تکرار هستند. و به ازای اجراهای متفاوت نتایج یکسانی تولید می‌کنند و امکان رسیدن به درک بهتری از تخمین تلاش پروژه را به ازای جابجایی فاکتورهای تعیین کننده فراهم می‌کنند. این دسته روش‌ها بسیار وابسته به مقادیر ورودی هستند. مقادیری که برای کالیبره کردن این الگوریتم‌ها بکار گرفته می‌شوند ممکن است در طی بکار گیری برای پروژه‌های جدید قابل استفاده نباشند. در ضمن بیشتر این الگوریتم‌ها وابسته به کمپانی توسعه دهنده خود هستند و در حالت کلی برای همه‌ی پروژه‌ها

[1] M. Khosravi is with Department of Computer Engineering, Islamic Azad University, Broujerd Branch, Broujerd, Iran (email: mehran_khosravy@yahoo.com)
L. Rikhtechi is with Department of Computer Engineering, Islamic Azad University, Broujerd Branch, Broujerd, Iran (email: rikhtechileila@iaub.ac.ir)
S. Andalibi is is with Department of Computer Engineering, Islamic Azad University, Broujerd Branch, Broujerd, Iran (email: shahramandalibi@yahoo.com).

نتایج مطلوبی تولید نمی‌کنند.

نمونه این روش‌ها، روش فاکتورهای هزینه[2]، مدل‌های خطی[3]، مدل‌های تجمعی[4]، مدل‌های توابع توانی[5]، رگرسیون خطی[6]، مدل‌های گسسته و مدل‌های اکتشافی می‌باشد.

۳.فاکتورهای تخمین تلاش

تمامی روش‌هایی که در مورد تخمین تلاش پروژه‌های نرم افزاری کار می‌کنند برای اینکه بتوانند نتایج یکسان و قابل مقایسه‌ای تولید کنند از تعداد مشخصی از فاکتورهای تخمین تلاش استفاده می‌کنند. این فاکتورها را در ۴ گروه اصلی طبقه بندی نموده‌اند که شامل ویژگی‌های محصول، ویژگی‌های بستر کامپیوتری، ویژگی‌های پرسنل و ویژگی‌های پروژه می‌شود. هر کدام از این گروه‌ها دارای زیر مجموعه‌ای از معیارهای مختلف می‌باشند که با مقادیر ۵ گانه خیلی کم، کم، نرمال، زیاد و خیلی زیاد مقداری دهی شده‌اند و هر معیار یکی از ویژگی‌های پروژه مورد بررسی را شامل می‌شود. مثلا برای ویژگی محصول معیارهای جدول(۱) وجود دارد.

جدول ۱: فاکتورهای ویژگی‌های محصول

فاکتور های هزینه		توصیف
		ویژگی های محصول
1	RELY	میزان قابلیت اعتماد لازم برای نرم‌افزار
2	DATA	اندازه پایگاه داده
3	CPLX	پیچیدگی محصول
4	DOCU	مستند سازی متناسب با نیازهای چرخه حیات
5	RUSE	توسعه با قابلیت استفاده مجدد
6	RCPX	قابلیت اطمینان محصول

۴.معیارهای ارزیابی

در زمینه تخمین تلاش پروژه‌های نرم افزار کارهای مختلف انجام شده است. برای آنکه بتوان کارهای مختلف را از نظر میزان کارایی با یکدیگر مقایسه نمود باید معیار یکسانی را تعریف نمود. در بیشتر مقالات از MRE(Mean Absolute Error) و RMSE(Root Mean Squared Error) نیز برای نشان دادن تفاوت عملکرد دو تکنیک متفاوت استفاده شده است. روابط (۱) و (۲) روش به دست آوردن این دو معیار است.

$$MRE=\frac{|a_1-c_1|+|a_2-c_2|+\ldots+|a_n-c_n|}{n} \quad (1)$$

$$RMSE=\sqrt{\frac{(a_1-c_1)^2+(a_2-c_2)^2+\ldots+(a_n-c_n)^2}{n}} \quad (2)$$

که در این دو رابطه a مقدار خروجی واقعی و c مقدار خروجی قابل قبول است. با این دو معیار می‌توان میزان خطای پیش بینی از مقدار واقعی را تعیین نمود و هر اندازه این مقدار کمتر باشد مدل توصیف شده عملکرد بهتری خواهد داشت. در کنار این دو معیار می‌توان از Pread نیز استفاده نمود. این معیار نشان دهنده تعداد پیش بینی‌هایی است که خطای آنها از مقدار مورد نظر کمتر است. مثلا Pread(30) نشان دهنده تعداد پیش بینی‌هایی است که میزان خطای آنها کمتر از ۳۰ واحد است. در ضمن برای اینکه بتوان نتایج تولید شده توسط روش پیشنهادی را با سایر روش‌ها مقایسه نمود باید از مجموعه داده‌های استانداردی بهره گرفت. به این منظور در زمینه پیش بینی تلاش بنچ مارک‌هایی موجود می‌باشد که معروف ترین و پرکاربرد ترین آنها را می‌توان از سایت ناسا[7] تهیه نمود. برای این مقاله ما از مجموعه داده‌های NASA93 و Cocomo81 استفاده کرده‌ایم. مجموعه داده‌های NASA93 شامل ۹۳ پروژه انجام شده در پایگاه ناسا می‌باشد که از ۱۷ فاکتور برای تعیین ویژگی‌های پروژه استفاده کرده است. مجموعه داده‌های Cocomo81 شامل ۶۳ پروژه می‌باشد که از ۱۴ فاکتور برای تعیین ویژگی‌های پروژه ها استفاده کرده است. تمامی مقالات ارائه شده در این زمینه از این دو پایگاه داده استفاده کرده‌اند.

۵.کارهای مرتبط

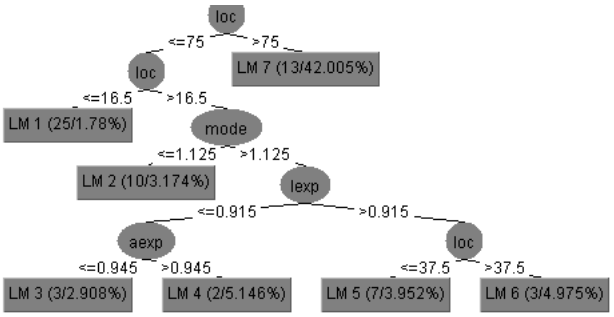
در این زمینه کارهای مختلفی صورت گرفته است که هریک نتایج خاص خود را داشته‌اند و متاسفانه روشی که قابلیت بکار گیری بر روی همه پروژه‌ها را داشته باشد معرفی نشده است. مثلا Hims & Mensies (۲۰۰۶) [۱] بر روی روش‌های مبتنی بر رگرسیون عمل کرده و بهترین MRE آنها ۴۰ و بهترین Pred آنها ۵۰ بوده است. Mahajon و همکارانش (۲۰۱۱) [۲] از آنالیز اکتشافی استفاده کرده‌اند و بهترین MRE آنها ۴۰ و بهترین Pred آنها ۵۰ بوده است. Attarzadeh و همکارانش (۲۰۱۱) [۳] از تکنیک‌های محاسبات نرم استفاده کرده‌اند که بهترین MRE آنها ۳۶ و Pred آنها ۵۰ بوده است.Kaur و همکارانش (۲۰۱۰) از تکنیک‌های آنالیز اکتشافی استفاده کرده‌اند که تنها میزان MRE را گزارش کرده‌اند که ۱۲ بوده است.

۶.کارهای مرتبط

با توجه به ویژگی‌های ذکر شده در بخش قبل در مورد درخت‌های تصمیم‌گیری در صورتیکه از این درخت‌ها برای داده کاوی مجموعه داده‌های موجود برای پیش بینی پرش استفاده کنیم، می‌توانیم به مدل‌هایی برسیم که نسبت به مدل‌های قبلی دارای تطبیق پذیری و جامعیت بیشتر بوده و امکان بکار گیری در پروژه‌های نرم افزاری مختلف را خواهند داشت. عملیات داده کاوی و ساخت مدل‌ها را بر روی هر دو مجموعه NASA93 و Cocomo81 انجام می‌دهیم و در هر بار از مجموعه‌های آموزشی بین ۴۰ تا ۸۰ رکوردی برای آموزش مدل‌ها استفاده می‌کنیم. عملیات تست را با مجموعه‌های ۱۰ رکودری که هر بار به صورت تصادفی از داده‌هایی که در آموزش شرکت نداشته‌اند انجام می‌دهیم. شکل ۱ نمونه درخت تولید شده از داده

[2] Cost Factors
[3] Liner Models
[4] Multiplicative Models
[5] Power Function Models
[6] Linear Regression
[7] http://mdp.ivv.nasa.gov/

کاوی بر روی مجموعه داده Cocomo81 است.



شکل ۱: درخت تولید شده از فرآیند داده کاوی بر روی مجموعه داده

همان گونه که از شکل ۱ هم مشخص است با اجرای فرآیند داده کاوی ۶ قانون از مجموعه داده‌ها استخراج می‌شود. یکی از ویژگی‌های بسیار مهم و جالب درخت‌های تصمیم گیری این است که این درخت ها تصمیم خود را شرح می دهند، بعبارت دیگر با دنبال کردن مسیر از نود ریشه به طرف برگ‌ها می‌توان دلیل تولید یک قانون خاص را مورد تحلیل و ارزیابی قرار داد، این امر امکان کالیبراسیون داده‌ها برای رسیدن به جواب مطلوب تر را فراهم می‌کند. ساختار و دلیل قوانین تولید شده بصورت زیر می‌باشد :

```
M5 pruned model tree:
(using smoothed linear models)
loc <= 75 :
|   loc <= 16.5 : LM1 (25/1.78%)
|   loc >  16.5 :
|   |   mode <= 1.125 : LM2 (10/3.174%)
|   |   mode >  1.125 :
|   |   |   lexp <= 0.915 :
|   |   |   |   aexp <= 0.945 : LM3 (3/2.908%)
|   |   |   |   aexp >  0.945 : LM4 (2/5.146%)
|   |   |   lexp >  0.915 :
|   |   |   |   loc <= 37.5 : LM5 (7/3.952%)
|   |   |   |   loc >  37.5 : LM6 (3/4.975%)
loc >  75 : LM7 (13/42.005%)
```

برای آنکه بتوان درک بتوان درک بهتری از قوانین تولید شده توسط این درخت پیدا کرد، نمونه‌ای از این قوانین در ادامه آورده شده است :

LM num: ۱

acteffort =

- ۴۱۰,۵۶۴۵mode
- ۵۵۴,۴۴۵۷ +time
- ۳۲۱,۴۹۴۳ -lexp
- ۳,۶۱۹۲ +loc

۶۹۷,۵۲۱۳ -

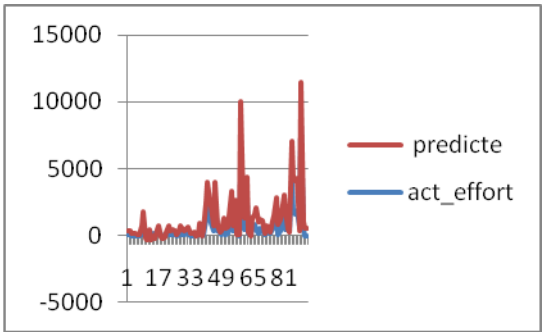
LM num: ۲

acteffort =

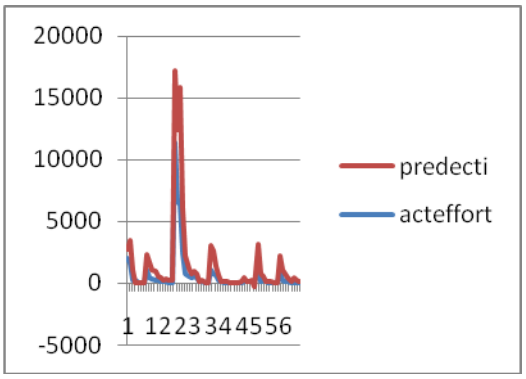
- ۱۰۱۵,۷۲۳۷mode
- ۵۵۴,۴۴۵۷ +time
- ۹۸۴,۶۲۹۴ -lexp
- ۵,۸۱۱۹ +loc

۷۴۵,۷۲۱۲ -

برای نشان دادن دقت این روش علاوه بر آنکه مقادیر Pred(30) MRE در ادامه آورده خواهد شد. نمونه‌ای از نمودار های رسم شده برای کل رکوردهای هر دو مجموعه داده های بکار گرفته شده در این روش در شکل ۲ و شکل ۳ نشان داده شده است.



شکل ۲: نمودار بر روی مجموعه داده های NASA۹۳



شکل ۳: نمودار بر روی مجموعه داده های Cocomo۸۱

همانوگونه که از روی هر دو نمودار مشخص است تطبیق مقادیر واقعی و مقادیر پیش بینی شده در سطح بسیار نزدیکی می‌باشد. و در دو نمودار علاوه بر اینکه الگوی حرکتی هر دو مقدار یکسان است، مقادیر نیز به یکدیگر بسیار نزدیک هستند. برای تکمیل این ارزیابی جدول مقادیر MRE و Pred(30) هر دو مجموعه داده در جدول های ۲ و ۳ آورده شده است.

۷.نتیجه گیری

با توجه به مقادیر ارائه شده در جدول‌های ۲ و ۳ مشاهده می‌شود که مقادیر MRE و Pred نسبت به روش‌های پیشین که در بخش ۲-۱ به آنها اشاره شده نتایج بهتری را تولید می کنند در ضمن با بررسی نمودارهای ارائه شده می‌توان نتیجه گرفت که مدل پیشنهادی می‌تواند الگوی رفتاری و تغییرات پروژه‌های مختلف را در ارزیابی خود دخالت دهد و این الگوها را به مدل نهایی اعمال کند.

جدول۲: مقادیر ارزیابی برای NASA۹۳

	واقعی	پیش بینی	MRE
۱	480	474.94	0.010
۲	1350	1349.4	0.001
۳	703	722.57	0.027
۴	48	38.217	0.203
۵	576	592.84	0.029
۶	882	894.29	0.013
۷	571.4	571.26	0.001
۸	42	44.435	0.057
۹	42	40.859	0.027
۱۰	60	56.353	0.060
MMRE= 4.28 – 127.56 PRED(30)= 45.16			

جدول۳: مقادیر ارزیابی برای Cocomo۸۱

	واقعی	پیش بینی	MRE
۱	106	69.61	0.343
۲	724	777.7	0.074
۳	82	83.67	0.020
۴	12	15.16	0.263
۵	41	43.7	0.065
۶	70	74.04	0.057
۷	55	40.16	0.269
۸	88	70.52	0.198
۹	230	207.3	0.093
۱۰	387	421.9	0.090
MMRE= 14.72 -- 172.44 PRED(30)= 26.98			

سپاسگزاری

از استاد بزرگوار جناب آقای دکتر سعید پارسا(دانشگاه علم و صنعت) که ما را در ارائه این مقاله یاری نمودند و مجموعه داده‌های آموزشی را در اختیار ما قرار دارند صمیمانه سپاسگذاریم.

REFERENCES

1 Hihn J, Menzies T. Anonymous. NASA Software Assurance Symposium, 2006.
2 Mahajan J, Devanand A, Dhruve K. Reusability Based Effort Estimatin Technique Using Dynamic Neural Network, 2011.
3 Attarzadeh I, Ow SH. A Novel Algorithmic Cost Estimation Model Based on Soft Computing Technique. Journal of Computer Science, 2010; 6 (2): 117-125.
4 Kaur J, Singh S, Singh Kahlon K, Bassi P. Neural Network: A Novel Technique for Software Effort Estimation. International Journal of Computer Theory and Engineering, 2010; 2(1): 1793-8201.