Computing and Informatics, Vol. 26, 2007, 1–15

AN ALGORITHM FOR THE GENERATION OF SEGMENTED PARAMETRIC SOFTWARE ESTIMATION MODELS AND ITS EMPIRICAL EVALUATION

Juan J. Cuadrado-Gallego, Miguel-Angel Sicilia

Computer Science Department, Polytechnic Building, University of Alcalá
Ctra. Barcelona km. 33.6, 288 71 Alcalá de Henares, Madrid, Spain
e-mail: {jjcg, msicilia}@uah.es

Manuscript received 21 April 2005; revised 4 July 2006
Communicated by Clark Thomborson

Abstract. Parametric software effort estimation techniques use mathematical cost-estimation relationships derived from historical project databases, usually obtained through standard curve regression techniques. Nonetheless, project databases – especially in the case of consortium-created compilations like the ISBSG –, collect highly heterogeneous data, coming from projects that diverge in size, process and personnel skills, among other factors. This results in that a single parametric model is seldom able to capture the diversity of the sources, in turn resulting in poor overall quality. Segmented parametric estimation models use local regression to derive one model per each segment of data with similar characteristics, improving the overall predictive quality of parametrics. Further, the process of obtaining segmented models can be expressed in the form of a generic algorithm that can be used to produce candidate models in an automated process of calibration from the project database at hand. This paper describes the rationale for such algorithmic scheme along with the empirical evaluation of a concrete version that uses the EM clustering algorithm combined with the common parametric exponential model of size-effort, and standard quality-of-adjustment criteria. Results point out to the adequacy of the technique as an extension of existing single-relation models.

Keywords: Parametric software estimation, software project databases, clustering algorithms, EM algorithm

REFERENCES

[1] BOEHM, B.—ABTS, C.—CHULANI, S.: Software Development Cost Estimation Approaches - a Survey. USC Center for Software Engineering Technical Report # USC-CSE-2000-505, 2000.

[2] BOEHM, B.—ABTS, C.—WINSOR BROWN, A.—CHULANI, S.—CLARK, B.—HOROWITZ, E.—MADACHY, R.—REIFER, D.—STEECE, B.: Software Cost Estimation with Cocomo II. Prentice Hall, 2000.

[3] BOETTICHER, G.: When Will it Be Done? The 300 Billion Dollar Question, Machine Learner Answers. IEEE Intelligent Systems, May/June 2003, pp. 2–4.

[4] CONTE, S. D.—DUNSMORE, H. E.—SHEN, V. Y.: Software Engineering Metrics and Models. Benjamin/Cummings, Menlo Park, CA, 1986.

[5] CUADRADO-GALLEGO, J. J.—SICILIA, M. A.—RODRÍGUEZ, D.—GARRE M.: An Empirical Study of Process-Related Attributes in Segmented Software Cost-Estimation Relationships. Journal of Systems and Software, Vol. 3, 2006, No. 79, pp. 351–361.

[6] DEMPSTER, A. P.—LAIRD, N. M.—RUBIN, D. B.: Maximum-Likelihood from Incomplete Data Via the em Algorithm. J. Royal Statist. Soc. Ser. B., Vol. 39, 1977.

[7] DICK, S.—MEEKS, A.—LAST, M.—BUNKE, H.—KANDEL, A.: Data Mining in Software Metrics Databases. Fuzzy Sets and Systems, Vol. 145, 2004, No. 1, pp. 81–110.

[8] DOLADO, J.: On the problem of the software cost function. Information and Software Technology, Vol. 43, 2001, No. 1, pp. 61–72.

[9] DREGER, J. B.: Function Point Analysis. Englewood Cliffs, NJ: Prentice Hall, 1989.

[10] GARMUS, D.—HERRON, D.: Function Point Analysis: Measurement Practices for Successful Software Projects, Addison-Wesley, 2000.

[11] GARRE, M.—CUADRADO, J. J.—SICILIA, M. A.: Recursive Segmentation of Software Projects or the Estimation of Development Effort. Proceedings of the V ADIS 2004 Workshop on Decision Support in Software Engineering, CEUR Workshop Proceedings, Vol. 120, available at `http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-120/`.

[12] GRAY, A.—MACDONELL, S. G.: Applications of Fuzzy Logic To Software Metric Models for Development Effort Estimation. Proceedings of the 1997 Annual Meeting of the North American Fuzzy Information Processing Society NAFIPS, IEEE, Syracuse NY, 1997, pp. 394–399.

[13] LUNG, C. H.—ZAMAN, M.—NANDI, A.: Applications of Clustering Techniques to Software Partitioning, Recovery and Restructuring. Journal of Systems and Software, Vol. 73, 2004, No. 2, pp. 227–244.

[14] NELDER, J. A.—MEAD, R.: Computer Journal. Vol. 7, 1965, pp. 308–313.

[15] NESMA: NESMA FPA Counting Practices Manual CPM 2.0, 1996.

[16] OLIGNY, S.—BOURQUE, P.—ABRAN, A.—FOURNIER, B.: Exploring the Relation Between Effort and Duration in Software Engineering Projects in World Computer Congress. Beijing, China, August 21–25, 2000, pp. 175–178.

[17] Parametric Estimating Initiative: Parametric estimating handbook. 2nd edition, 1999.

[18] STENSRUD, E.—FOSS, T.—KITCHENHAM, B.—MYRTVEIT, I.: An Empirical Validation of the Relationship Between the Magnitude of Relative Error and Project Size. In Proceedings of the Eighth IEEE Symposium on Software Metrics, 2002.

[19] PEDRYCZ, W., SUCCI, G.: Genetic granular classifiers in modeling software quality. The Journal of Systems and Software, Vol. 76, 2002, pp. 277–285.

[20] WITTEN, I. H.—FRANK, E.: Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers, San Francisco, California, 1999.

[21] XU, Z.—KHOSHGOFTAAR, T.: Identification of Fuzzy Models of Software Cost Estimation. Fuzzy Sets and Systems, Vol. 145, 2004, No. 1, pp. 141–163.

Juan J. CUADRADO-GALLEGO is currently a Profesor Titular de Universidad at the Computer Science Department of the University of Alcalá, Madrid, Spain, and the Director of the doctorate studies in computer sciences at this University. He also is a consultant at the University Oberta of Catalunya, Barcelona, Spain. His expertise area is software engineering and especially software measurement. He has been teaching in this subject in all the universities where he has been as permanent staff and at the University Roma Tre, Roma, Italy, as teaching staff.

Miguel-Angel SICILIA obtained his university degree in computer science from the Pontifical University of Salamanca in Madrid, Spain (1996) and his Ph. D. from Carlos III University in Madrid, Spain (2002). In 1997 he joined an object-technology consulting firm, after enjoying a research grant at the Instituto de Automatica Industrial (Spanish Research Council). From 1997 to 1999 he worked as Assistant Professor at the Pontifical University. Since 2000 to October 2003, he worked as a full-time lecturer at Carlos III University working actively in the area of adaptive hypermedia and e-learning systems. Currently, he works as an Associate Professor at the Computer Science Department, University of Alcalá (Madrid). His research interests focus primarily on semantic metadata, software engineering and learning technology.