Warsaw School of Economics
Collegium of Economic Analysis
Institute of Information Systems and Digital Economy

MSc Przemysław Pospieszny

Doctoral Thesis Summary

Application of data mining techniques for effort and duration estimation of software projects

Written under supervision of
Professor Andrzej Kobyliński

Warsaw 2015

1. Introduction

This dissertation proposes an alternative approach to estimation of effort and duration of information and communication technology (ICT) projects in relation to traditional estimation methods based largely on expert knowledge, source lines of code and function points. For this purpose, predictive data mining techniques derived from statistics, machine learning and artificial intelligence have been applied. Prediction capability of data exploration algorithms are commonly acknowledged, which manifests itself in an increasingly larger scale of their application in different sectors of economy during the last two decades. They have specific use in the fields characterized by great complexity and uncertainty regarding final product or outcome such as: credit risk assessment, customer relationship management or fraud detection.

The concept of applying data exploration techniques to the estimation of effort and duration of software projects originates from a problem of a large number of initiatives (according to some studies: 65%[1]) that do not stick to costs and completion time agreed during the initiation and planning phase. As a result, products created in software projects have often different functional scope and low quality, which leads to dissatisfaction of the final customer. Additionally, higher than the assumed budget and duration necessary for product creation may result in unfavourable business profits balance in relation to incurred costs, leading to premature resignation from execution of the already started projects.

It is common to look for reasons for abandoning projects not at the stage of errors committed during creation of the final product, tests or implementation, but in the initiation and planning phase[2]. The aforementioned project assumptions are specified at the beginning of an initiative and rarely changed in later phases. At the same time, it is required to conduct the process of change request, which involves revision of benefit calculation. In the case of costs that exceed potential gains from implementation of the final product of a project, requested initiative

[1] Standish Group, *The CHAOS Manifesto 2011*, „The Standish Group International. EUA", 2011; B. Czarnacka-Chrobot, *Analysis of the functional size measurement methods usage by Polish business software systems providers*, „Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)", 2009, t. 5891 LNCS, pp. 17–34.
[2] G. Wells, *Why Projects Fail*, „Management Science Journal," 2003.

schedule or budget changes may not obtain approval of the sponsor and stakeholders and hence result in the project failure.

The estimation of project assumptions, labour consumption and duration in the early phases of project implementation is an extremely difficult task owing to incomplete available knowledge regarding the final product, initiative and tasks related to its implementation. Errors made in the process of estimation have a direct impact on the course of the project and its successful completion. However, in organizations it is frequent to estimate project assumptions on the basis of guidelines of managerial staff or customer. Additionally, the estimation process is based on traditional methods, such as estimation by analogy, expert methods or decomposition. In the case of inexperienced project managers, such activities may bring over optimistic assumptions with regard to effort and duration of initiatives, which can lead to failure in the further project phases, especially when the costs exceed benefits from ICT system implementation.

Organizations experienced in project management perceive accurate estimation as a prerequisite of successful completion of an initiative, and use advanced estimation methods based on dimensioning of software in the form of source line of code or function points. These approaches have been developed since the end of the 1970s and provide standardization, repeatability and continuous improvement of the project estimation process by updating and calibration of particular techniques. However, both approaches have many flaws limiting their common use in practice. Techniques based on SLOC are not adjusted to the contemporary programming languages and do not take into account labour consumption other than software development, such as requirements collection and tests[3]. At the same time, those based on function points require final product specification, thus they may be used in the initial project phase only in a limited form. Additionally, they are based on a subjective evaluation of the estimator[4]; system size, labour consumption and duration necessary for its manufacturing can be thus differently estimated as a result of personal evaluation of the assessor. The use of both SLOC and FP requires trained personnel who largely performs manual calculations both for the

[3] D. Galorath, M. Evans, *Software Sizing, Estimation, and Risk Management*, Auerbach Publications, Boca Raton 2006, p.12.
[4] C. Kemerer, *Reliability of function points measurement: a field experiment*, „Communications of the ACM," 1993, t.36, no 2, pp. 85–97.

purposes of dimensioning software and defining project parameters, which may result in frequently incorrect and over optimistic estimations in relation to the actual values.

For the aforementioned reasons, researchers involved in the issues of estimation of software projects[5] in recent years have turned their attention towards a new discipline involved in the exploration of large data sets: data mining. This concept was created in the early 1990s and had its source in the popularization of data warehouses, business intelligence and knowledge management. Due to combining techniques originating from different science disciplines, such as statistics, mathematics, artificial intelligence or machine learning, the algorithms of data mining are characterized by a great accuracy of estimation and have practical application in creating models that increase the economic value of an organization. Research works conducted during last 20 years proved their effective application with regard to budget estimation, schedule and final product quality (in relation to the number and type of errors)[6]. Additionally, the algorithms of data mining, of both supervised learning and non-supervised learning, may be used as a tool for monitoring progress in project implementation, for instance in the analysis Earned Value Management (EVM)[7] or for the prediction of future system maintenance costs[8].

Significant emphasis is put on estimating effort and duration of initiatives aimed at developing new or expanding the existing system at the initial stage of the project[9]. It is the largest challenge to an estimator due to incomplete information with regard to requirements related to the final product, uncertainty about activities related to its creation and substantial probability of risks

[5] D. Dzega, W. Pietruszkiewicz, *Classification and metaclassification in large scale data mining application for estimation of software projects*, „2010 IEEE 9th International Conference on Cybernetic Intelligent Systems, CIS 2010", 2010; K. Dejaeger et al., *Data mining techniques for software effort estimation: A comparative study*, „IEEE Transactions on Software Engineering", 2012, t. 38, pp. 375–397; I.F. de Barcelos Tronto, J.D.S. da Silva, N. Sant'Anna, *Comparison of Artificial Neural Network and Regression Models in Software Effort Estimation*, „Neural Networks, 2007. IJCNN 2007. International Joint Conference on", 2007, pp. 771–776.

[6] N.K. Nagwani, A. Bhansali, *A data mining model to predict software bug complexity using bug estimation and clustering*, „ITC 2010 - 2010 International Conference on Recent Trends in Information, Telecommunication, and Computing," 2010, pp. 13–17.

[7] S.H. Iranmanesh, Z. Mokhtari, *Application of data mining tools to predicate completion time of a project*, „Proceeding of World Academy of Science, Engineering and Technology," 2008, t.32, pp. 234–240.

[8] R. Shukla, A.K. Misra, *Estimating software maintenance effort a neural network approach*, „Proceedings of the 2008 1st India Software Engineering Conference, ISEC'08," 2008, pp. 107–112.

[9] C. Lopez-Martin, C. Isaza, A. Chavoya, *Software development effort prediction of industrial projects applying a general regression neural network*, „Empirical Software Engineering", 2012, t. 17, s. 738–756; J. Villanueva-Balsera et al., *Effort estimation in information systems projects using data mining techniques*, „Proceedings of the 13th WSEAS International Conference on Computers – Held as part of the 13th WSEAS CSCC Multiconference", 2009, pp. 652–657.

occurrence. For that purpose, researchers used individual prediction algorithms, which generated estimations of effort and duration of initiatives, and thus were showing their suitability for presenting the studied problem.

However, so far data exploration techniques were not commonly applied in practice in organizations implementing ICT systems as a tool supporting the process of estimation of resources necessary for the development of final products. This phenomenon may result from inconsistencies of research works conducted in this respect. The obtained results differed depending on techniques used, their configuration and using historical base of projects in the process of algorithms learning. Models applied were predominantly individual. This type of models implemented in practice can generate results different from those obtained in research works due to the diversity of data in a selected organization. Additionally, most models were built with use of data sets smaller than 100 observations and coming from one selected institution. Hence there is a significant possibility of overfitting and distorting actual predictive power of algorithms. Another aspect is quality of projects data, which is low in many organizations. Meanwhile, algorithms proposed by the researchers should be resistant to missing values and noises in data. The above inconsistencies in research works, lack of proposal of integrated approach towards estimation of IT projects and unpopularity of using these methods in practice contributed to addressing these drawbacks in this doctoral dissertation.

2. Subject of the dissertation and research methodology

The aim of the dissertation is application of ensemble predictive data mining techniques for the estimation of effort and duration of software projects at their initial stage for the purpose of models' development that could be utilitized in practice. In order to conduct the research, ISBSG[10] database was used, containing historical data about the completed ICT initiatives from many public and private institutions, operating in different branches of industry, concerning both developing new software and enhancing the existing one. Construction of the abovementioned data exploration models was conducted according to recognized methodology

[10] International Software Benchmarking Standards Group, *ISBSG Repository Data Release 12 - Field Descriptions*, 2013.

of *Cross Industry Standard Process for Data Mining* (CRISP-DM)[11]. First, data analysis was performed for the purpose of selection of models' training data set, identification of associations between variables and their impact on dependent variables (effort and duration). Three prediction data exploration algorithms were chosen for models' construction: generalized linear models (GLM), artificial decision trees (ANN) and CHi-squared Automatic Interaction Detection decision trees (CHAID). The choice of these algorithms was made on the basis of review of the literature and results of preliminary analysis of models that preferably corresponded to the examined phenomena and were characterized by resistance to missing and noisy data. These techniques were used for construction of two ensemble models, separately for dependent variables: effort and duration. Each of models consists of three abovementioned algorithms, whose results were averaged, in accordance with the best practice, for the purpose of a more accurate estimation. The model evaluation was supposed to confirm the ability to predict the studied phenomena, greater accuracy of estimation of aggregation model than individual techniques as well as the possibility of implementation of models in practice.

The approach presented above enables estimating software projects related not only to development of new or enhancement of the existing system in practice, but also to implementation or change in the methodology of work, processes and procedures implemented within the scope of ICT project management environment. Ensemble techniques enable organizations to efficiently implement the models in practice, which consists in preparing data and adjusting generated for the purposes of this dissertation models to reflect diversity of initiatives within a given institution. As a result of combining predictive power of three effective algorithms, the intention of the author was to create a tool resistant to poor quality of input data and uniqueness of particular organisations in terms of work culture and project data. Models were constructed with the use of IBM SPSS Modeler, which is commonly used by institutions using data exploration techniques for various day-to-day operations. The additional outcome of this dissertation is formulation of models' implementation methodology proposal.

[11] C. Pete et al., *CRISP-DM 1.0*, *CRISP-DM Consortium*, 2000.

3. Objectives and research hypotheses

The main objective of dissertation was construction of aggregating prediction models with the use of general linear models, multilayer perceptron artificial neural networks and CHAID decision trees for estimation of effort and duration of information technology projects. Below secondary research objectives were presented, enabling implementation of assumed main objective and formulated hypotheses.

Cognitive objectives:

- Determination of relations between variables describing software projects and their impact on estimation of effort and duration of initiatives.
- Assessment of suitability of general linear models, artificial neural networks and CHAID decision trees for the estimation of effort and duration of software projects.

Methodical objectives:

- Definition of approach for building ensemble prediction data mining models for estimation of effort and duration of IT projects with the use of three data mining regression techniques: general linear models, artificial neural networks and CHAID decision trees.
- Proposing methodology of implementation of the built models in practice.

Application objective:

- Construction of ensemble models estimating effort and duration of initiatives with the use of a multi-trade historical base of software projects for the purpose of initial calibration of prediction algorithms and assessment of their application possibilities, as a result of potential implementation of a model within the scope of management of initiatives in various types of organizations implementing software projects.

In this dissertation, the following research hypotheses were formulated:

1. Data mining can be applied in software project management, supporting the process of estimation of effort and duration of initiatives at their initial stage and potentially can contribute to increased probability of successful project completion. For these reason, they may be applied as an alternative to traditional methods and to methods using source lines of code or function points.

2. General linear models, artificial neural netowrks and CHAID decision trees are characterized by a sufficiently good prediction power of effort and duration of software projects and resistance to deficiencies and noises in data enabling potential implementation in practice.
3. Ensemble prediction data exploration models, used for estimation of software projects at the initial stage, enable obtaining more detailed estimates regarding studied phenomena than individually used algorithms.

4. Content and layout of the dissertation

The doctoral dissertation was divided into four chapters. In chapter one factors affecting ICT projects along with criteria of their successful completion were discussed. They are the starting point of understanding the issues of estimation of project parameters because their correct estimation affects the possibility of creating project final product in accordance with the predefined effort and duration. In subsequent subchapters issues related to estimation projects and measures used in the estimation process were discussed. Traditional techniques of estimation have been presented as well. They include: use of analogy, expert's estimation or decomposition, as well as more advanced parametric methods based on source line of code (COCOMO/COCOMO II, SLIM, SEER-SEM) and on dimensioning of software with the use of function points (IFPUG, NESMA, COSMIC). This chapter discusses also imperfections of the currently used techniques of estimating information technology projects.

Chapter two discusses knowledge discovery in project management and data exploration techniques. It covers also the review of literature related to applying data mining techniques for the estimation of project parameters. The areas of knowledge distinguished in the project management process, information collected at each stage in the form of data sets, containing such projects characteristics as budget, effort, duration, used programming language, methodology of work implementation or the number of errors detected during the tests phase (quality) were presented in the first place. Then, the process of knowledge discovery in data sets was discussed, which includes an important step: data mining. In particular subchapters types of techniques data mining were presented, methodology of process of knowledge acquisition from CRISP-DM data and three prediction algorithms, which in chapters 3 and 4

were used for construction of models for estimating effort and duration of projects: general linear models, artificial neural networks and CHAID decision trees. Further part of the chapter contains a review of previous literature on the use of exploration data techniques for the estimation of IT project parameters, indicating used approach, techniques, bases used for the learning process and methods of evaluation of models constructed with data mining algorithms. An important aspect of this part of the dissertation is the discussion on limits of past research from the area of the application of prediction models.

Chapter three initiates the empirical part of the dissertation, where CRISP-DM methodology was utilitized and ISBSG database was applied for construction of effort and duration prediction models. The first step intended to prepare a set of input data for prediction modelling of labour consumption and duration was understanding and analysis of data with regard to occurring missing values, conducting data transformation and analysis of Pearson correlation and stepwise regression. Then, separately for effort and duration of initiatives, process of construction and configuration of aggregation models was presented, consisting of three prediction techniques: generalized linear models, artificial neural networks and CHAID decision trees, along with substantiation for their choice.

Chapter four is an evaluation of individual models and ensemble ones in terms of estimation accuracy and possibility of using them in practice. Traditional methods of forecasting error were chosen, such as: average error, mean absolute error, mean squared error and root mean squared error and commonly used for evaluation of models labour consumption and duration of IT projects: mean relative error (MRE), mean magnitude of relative error (MMRE) and prediction to actual value ratio (PRED). First, each of three models constructed separately for estimating the effort and duration with the use of general linear models, artificial decision trees and CHi-squared Automatic Interaction Detection decision trees CHAID was assessed. Secondly, two aggregating models, each consisting of three aforementioned algorithms estimating researched phenomena, were subjected to evaluation. Results of an integrated approach were compared with the applied individual models for the purpose of ascertaining superiority manifesting itself in accuracy of estimated values. The last part of chapter four presents a proposal of methodology for implementation of aggregation model in practice.

The dissertation is summed up with conclusions from the conducted research and reference to assumed research objectives and research hypotheses. Additionally, thesis limitations are also presented, as well as potential directions of future research.

In the last part of the dissertation reference literature and appendices presenting the detailed results of built prediction models were listed.

5. Results and conclusions from the conducted research

The primary objective of the dissertation was to build two ensemble prediction models, separately for effort and duration estimation of software projects, with utilization of three data mining techniques: generalized linear model, multilayer perceptron artificial neural network and CHAID decision trees. A multi-trade historical base of software projects ISBSG, containing a large volume of data of good quality, was chosen for the process of training and validation of models. It is used by ISBSG member organizations to support process of new initiatives' sizing and estimation. The aforementioned approach enabled building models that are characterized by a very high accuracy of effort and duration estimations of projects. The obtained error measures, presented in the table 1 and 2, both for individual algorithms and for ensemble models were low, at the same time conforming to the Conte criterion[12] (mean magnitude of relative error MMRE lower than 0.25). Nonetheless, the applied ensemble approach based on averaging the results of individual techniques for effort and duration prediction generated more precise estimation than individual algorithms.

Table 1 Comparison of individual and ensemble models for effort estimation

	Generalized linear model		Multilayer perceptron ANN		CHAID decision tree		Ensemble model	
	Training	Test	Training	Test	Training	Test	Training	Test
Min error	-1,523	-1,170	-1,507	-1,410	-1,528	-1,334	-1,389	-1,227
Max error	1,172	1,073	1,334	1,192	1,362	1,200	1,230	1,155
ME	0,000	-0,012	0,008	0,002	0,000	-0,008	-0,004	-0,011
MAE	0,288	0,310	0,308	0,331	0,287	0,313	0,288	0,310

[12] S.D. Conte, H.E. Dunsmore, and V.Y. Shen, *Software engineering metrics and models*, Benjamin/Cummings Pub. Co. 1986.

MSE	0,139	0,162	0,159	0,175	0,140	0,169	0,139	0,160
RMSE	0,373	0,402	0,398	0,418	0,374	0,412	0,373	0,400
MMRE	0,203	0,053	0,226	0,113	0,225	0,050	0,187	0,040
PRED(0,25)	0,599	0,604	0,571	0,545	0,612	0,607	0,618	0,597
PRED(0,3)	0,680	0,662	0,657	0,623	0,680	0,662	0,685	0,662
Standard deviation	0,373	0,403	0,398	0,419	0,374	0,412	0,368	0,397
Linear correlation	0,768	0,713	0,729	0,687	0,767	0,698	0,775	0,722

Source: Own elaboration

Table 1 presents main evaluation indicators for three individual algorithms and ensemble model for prediction of man-months necessary for project's final product development. According to the data included, the aggregation model generates forecast errors at the level comparable to the best algorithm (generalized linear model). However, while analysing the measures of adjustment to data and estimation capability, ensemble model had the lowest forecasting error (MMRE): 0.187 for training set and 0.04 for test set. The indicator is on good level, proving high quality of the model and its capability for effort prediction. Additionally, the indicator PRED (0.25) for the aggregation model was also at the best level with regard to individual models and it amounted to approximately 60% for both used data sets (training and test dataset).

Table 2 Comparison of individual and ensemble models for duration estimation

	Generalized linear model		Multilayer perceptron ANN		CHAID decision tree		Ensemble model	
	Training	**Test**	**Training**	**Test**	**Training**	**Test**	**Training**	**Test**
Min error	-1,511	-1,085	-1,393	-1,048	-1,452	-0,964	-1,411	-1,032
Max error	0,845	0,702	0,904	0,660	0,971	1,082	0,907	0,804
ME	0,000	0,003	0,000	0,009	0,000	0,012	0,000	0,008
MAE	0,206	0,217	0,188	0,198	0,193	0,212	0,186	0,201
MSE	0,075	0,072	0,065	0,063	0,068	0,074	0,064	0,064
RMSE	0,274	0,268	0,255	0,250	0,261	0,273	0,253	0,252
MMRE	0,228	0,263	0,213	0,259	0,217	0,251	0,205	0,245
PRED(0,25)	0,611	0,558	0,623	0,588	0,654	0,568	0,659	0,591
PRED(0,3)	0,700	0,646	0,706	0,653	0,732	0,653	0,750	0,675
Standard deviation	0,275	0,268	0,255	0,250	0,261	0,273	0,253	0,253
Linear correlation	0,660	0,658	0,715	0,712	0,700	0,648	0,722	0,706

Source: Own elaboration

Similarly to effort prediction, the aggregation model for duration estimation (table 2) generates better forecasts than individual techniques. Error measures of this model (ME, MAE and RMSE) are close to the results of the artificial neural network, which is the most effective individual algorithm, generating slightly better estimates than the other two techniques. Mean

magnitude of relative error (MMRE) obtained for the ensemble model was on level of 0.205 (training) and 0.245 (testing). Both values indicate that aggregating model can generate approximately 20-24% wrong estimates. MMRE at the aforementioned level suggests very good capacity of the model to generate accurate predictions of months necessary for implementation of the project. PRED (0.25) values amount to, respectively for the teaching set: 66% and for the training set: 60%, that is as well on very good level.

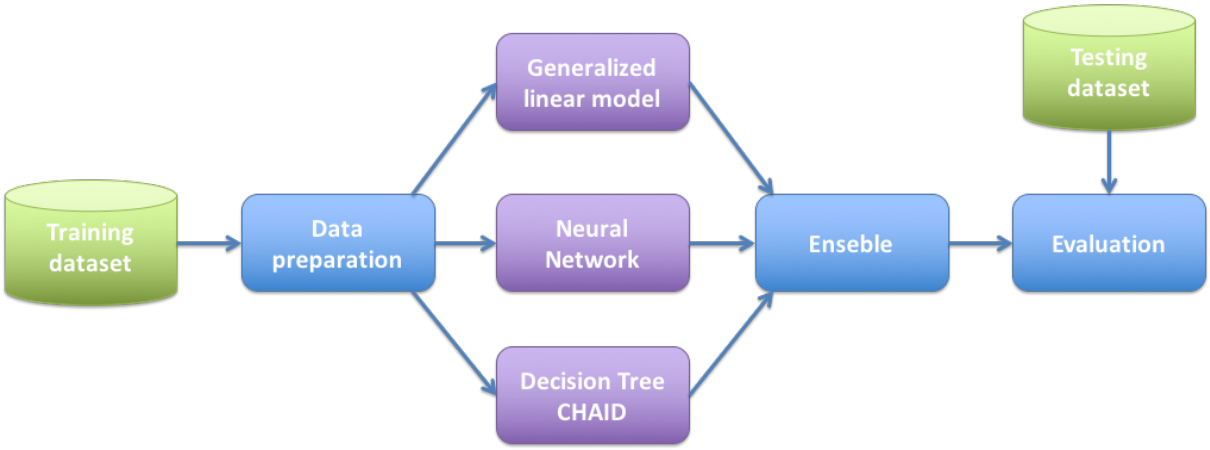

Figure 1 Process of building ensemble models

Source: Own elaboration

By implementation of mentioned above main objective of this dissertation secondary objectives were achieved, foremost the application objective. Its essence was preparation of data mining models for estimation of projects parameters, ensuring their accurate estimation, regardless of the type of IT initiative and organization in which it would be deployed. This objective was achieved as a result of applying an approach of aggregating three data mining prediction algorithms derived from techniques based on regression and machine learning: general linear models, artificial decision trees and CHAID decision trees (fig. 1). According to the obtained results (table 1 and 2), combining algorithms by averaging results obtained increased accuracy of estimation effort and duration estimation and potentially increased robustness of the models to noises in data and as well as to non-standard values and outliers. Additionally, it reduced the possibility of overfitting occurrence of a particular algorithm to data, which could result in misleading predictions. At the same time, using for training and validation the ISBSG database, which provides information about projects coming from different kinds of organizations and types of initiatives, enabled their initial calibration for potential implementation in any institution implementing IT projects. It is very important factor, since every organization has

different culture of work, project management methodology and as level of employee competence. Therefore, both effort and duration related to carrying out initiatives may differ significantly between organizations while implementing the same type of project.

Figure 2 Proposed methodology for implementation of ensemble models in practice based on CRISP-DM

Source: Own elaboration

Within the scope of the prepared methodical objectives, the developed methodology of implementation of obtained models in organizations interested in using them for supporting decision-making process of estimation of initiatives was presented (fig. 2). The suggested approach is an extension of CRISP-DM methodology by additional post-implementation steps that enable proper use of models by integrating them with the existing processes and tools used in a given organization, monitoring their performance and constant update. It requires availability of high quality database of historical ICT projects distinctive for particular institution, that enable adjusting models to the individual nature of initiatives implemented within the institution.

The consequence of models construction was the achievement of cognitive objectives of the dissertation, namely identification of the relation between variables, their impact on the prediction of effort and duration, and assessment of suitability of three selected algorithms for project parameters estimation. ISBSG projects database underwent the process of data preparation, which involved an analysis of Spearman and Pearson correlation and stepwise regression. The size of built final product had the greatest impact on effort of initiatives. Other input variables lesser impact on predicted variables, however, their influence was significant. In terms of duration estimation, interaction between dependent variable and independent variables was more evenly spread, with indication to the software size, the applied methodology, type of the system hardware platform and the required level of adjustment of a system to business requirements.

The implementation of the main objective and supporting ones (application, methodical and cognitive objectives) enabled verification and confirmation of three research hypotheses defined in this dissertation. According to the results presented in table 1 and 2, the prediction data mining techniques can be used in project management to estimate effort and duration of initiatives. Certainly, they may be used as a tool supporting decision-making and supplement traditional and parametrical estimation techniques. They enable more precise formulation of project assumptions at the initial stages, which increases the probability of successful initiative completion. Generalized linear model, artificial neural network and CHAID decision trees are characterized with a very high prediction power regarding the studied phenomena and their aggregation enables obtaining more accurate estimates and counteracts the possibility of overfitting occurrence of a given algorithm to the data.

The previous studies related to application of data mining techniques for software projects estimation were largely devoted to the issue of effectiveness of particular algorithms, using at the same time 20-30 years old databases with a small number of observations. Additionally, results presented in the studies were often inconsistent owing to the application of various approaches to preparation of data and the process of their construction. Therefore, it hardly can be found up to date any their application by organizations implementing software projects. As a result, the approach proposed in this dissertation, based on combining three effective data exploration algorithms, their initial calibration on the basis of a multi-trade base of historical projects, and suggested methodology for their implementation and maintenance may enable

their easier implementation in practice. The constructed models for prediction of effort and duration prediction are an alternative or supplementing method of estimation to traditional methods or methods based on function points and source line of code. As opposed to available approaches, they are an automatic tool for decision support delivering accurate estimates of initiatives parameters that does not require a considerable labour consumption, making estimation comes down to including in models characteristics of a new, previously unknown initiative. An additional advantage of models is ease of their implementation and future maintenance. They require only adding newly finished initiatives for the purpose of adjusting models to the rapidly changing nature of conducting IT projects in an organization on order to achieve constant accurate effort and duration estimates.

While summing up the dissertation, it is useful to mention its limitations, which may be eliminated as a result of future research. Both for construction and evaluation of models, one ISBSG database was used. It was divided into a training and a test dataset. The reason for this was lack of another reliable available data set that could be used for the verification of accuracy of estimates achieved with the selected algorithms. A preferred solution would be implementation of created models for estimation of effort and duration of projects in one selected organization or multiple organizations, where their ability to represent studied phenomena could be confirmed in practice. In this way, the suggested implementation-maintenance methodology could be verified. Another limitation of this dissertation, which may also constitute a basis for subsequent research, is the way of determining discrete variable representing size of generated product, which to the greatest extent influences the studied phenomena in comparison to other data within the data set. Its values in ISBSG database are based on intervals calculated with FSM methods. For the purpose of determination of this value, traditional techniques of data estimation may also be used, such as techniques based on expert judgement or analogy. However, the method of function points is commonly considered as the most precise technique of determination of software size. For this reason, future research could involve development of a detailed approach to software project parameters estimation using function points for the purpose of determination the size of the system and proposed in this dissertation approach of ensemble data mining models for effort and duration estimation necessary for project implementation. Additionally, influence of the size of FSM calculated with various methods on quality of the obtained predictions with *data mining* techniques, relating to effort and duration, could be compared for the purpose of selection of the most

adequate FSM method which, in combination with data exploration techniques, provides the most precise estimation of mentioned above project parameters.

Selected bibliography positions

Albrecht A.J., Gaffney J.E.. J., *Software Function, Source Lines of Code, and Development Effort Prediction: A Software Science Validation*, „IEEE Transactions on Software Engineering," 1983, t.SE-9.

Azzeh M., Cowling P.I., Neagu D., *Software stage-effort estimation based on association rule mining and Fuzzy set theory*, „Proceedings - 10th IEEE International Conference on Computer and Information Technology, CIT-2010, 7th IEEE International Conference on Embedded Software and Systems, ICESS-2010, ScalCom-2010," 2010, pp. 249–256.

Balsera J.V., Montequin V.R., Fernandez F.O., González-Fanjul C.A., *Data Mining Applied to the Improvement of Project Management*, „InTech," 2012.

De Barcelos Tronto I.F., da Silva J.D.S., Sant'Anna N., *Comparison of Artificial Neural Network and Regression Models in Software Effort Estimation*, „Neural Networks, 2007. IJCNN 2007. International Joint Conference on," 2007, pp. 771–776.

Boehm B.W., *Software Engineering Economics*, „Prentice Hall.," 1981, t. 10, pp. 4 –21.

Cios K., Pedrycz W., Swiniarski R., Kurgan L., *Data Mining A Knowledge Discovery Apporach*, Springer Science, New York, New York, USA 2007.

Clarke B., Fokoue E., Zhang H.H., *Principles and Theory for Data Mining and Machine Learning*, Springer Science, New York, New York, USA 2009.

Conte S.D., Dunsmore H.E., Shen V.Y., *Software engineering metrics and models*, Benjamin/Cummings Pub. Co. 1986.

Czarnacka-Chrobot B., *Analysis of the functional size measurement methods usage by Polish business software systems providers*, „Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)," 2009, t.5891 LNCS, pp. 17–34.

Czarnacka-Chrobot B., *Effectiveness of Business Software Systems Development and Enhancement Projects versus Work Effort Estimation Methods*, „International Journal of Social, Management, Economics and Business Engineering," 2013, t.7, nr 9, pp. 1329–1336.

Dejaeger K., Verbeke W., Martens D., Baesens B., *Data mining techniques for software effort estimation: A comparative study*, „IEEE Transactions on Software Engineering," 2012, t.38, pp. 375–397.

Dzega D., Pietruszkiewicz W., *Classification and metaclassification in large scale data mining application for estimation of software projects*, „2010 IEEE 9th International Conference on Cybernetic Intelligent Systems, CIS 2010," 2010.

Fayyad U., Piatetsky-Shapiro G., Smyth P., *From data mining to knowledge discovery in databases*, „AI magazine," 1996, pp. 37–54.

Flasiński M., *Zarządzanie projektami informatycznymi*, Wydawnictwo Naukowe PWN 2013.

Galorath D., Evans M., *Software Sizing, Estimation, and Risk Management*, Auerbach Publications 2006.

Gasik S., *A model of project knowledge management*, „Project Management Journal," 2011, t.42, nr 3, pp. 23–44.

Gatnar E., *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa 2008.
Giudici P., Figini S., *Applied Data Mining for Business and Industry*, John Wiley & Sons 2009.
Han J., Kamber M., Pei J., *Data Mining: Concepts and Techniques*, Morgan Kaufmann 2006.
Hand D.J., Mannila H., Smyth P., *Eksploracja danych*, Wydawnictwa Naukowo-Techniczne 2005.
Hill P., *Practical Software Project Estimation: A Toolkit for Estimating Software Development Effort & Duration*, McGraw Hill Professional 2010.
Iranmanesh S.H., Mokhtari Z., *Application of data mining tools to predicate completion time of a project*, „Proceeding of world academy of science, engineering and technology," 2008, t.32, pp. 234–240.
Jaszkiewicz A., *Inżyneria oprogramowania*, Helion, Gliwice 1997.
Jorgensen M., Shepperd M., *A Systematic Review of Software Development Cost Estimation Studies*, „IEEE Transactions on Software Engineering," 2007, t.33, nr 1, pp. 33–53.
Kemerer C., *Reliability of function points measurement: a field experiment*, „Communications of the ACM," 1993, t.36, nr 2, pp. 85–97.
Kisielnicki J., *Zarządzanie wiedzą we współczesnych organizacjach*, Wyższa Szkoła Handlu i Prawa im. Ryszarda Łazarskiego 2003.
Kisielnicki J., Zarządzanie projektami, Wydawnictwo JAK, Warszawa 2011.
Kobyliński A., *Miary procesu i produktu programowego*, „Współczesne kierunki rozwoju informatyki (PTI)," 2004, pp. 105–109.
Kobyliński A., Pospieszny P., *Zastosowanie technik eksploracji danych do estymacji pracochłonności projektów informatycznych*, „Studia i Materiały Polskiego Stowarzyszenia Zarządzania Wiedzą," Bydgoszcz 2015, pp. 67–82.
Laird L.M., Brennan M.C., *Software Measurement and Estimation: A Practical Approach*, John Wiley & Sons 2006.
Larose D.T., *Data Mining Methods and Models*, John Wiley & Sons 2007.
Linoff G.S., Berry M.J.A., *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, John Wiley & Sons 2011.
Lopez-Martin C., Isaza C., Chavoya A., *Software development effort prediction of industrial projects applying a general regression neural network*, „Empirical Software Engineering," 2012, t.17, pp. 738–756.
Marchewka J., *Information Technology Project Managment - Providing Measurable Organizational Value*, *Management*, Wiley 2003.
McConnell S., *Software Estimation: Demystifying the Black Art: Demystifying the Black Art*, Microsoft Press 2009.
Kearns M, Vailiant L., *Cryptographic Limitations on Learning Boolean Formulae and Finite Automata*, „Symposium on Theory of computing (ACM)," 1989, pp. 433–444.
Mittas N., Angelis L., *Ranking and clustering software cost estimation models through a multiple comparisons algorithm*, „IEEE Transactions on Software Engineering," 2013, t.39, nr 4, pp. 537–551.
Nagwani N.K., Bhansali A., *A data mining model to predict software bug complexity using bug estimation and clustering*, „ITC 2010 - 2010 International Conference on Recent Trends in Information, Telecommunication, and Computing," 2010, pp. 13–17.
Neimat T. Al, *Why IT projects fail*, „The project perfect white paper collection," 2005, pp. 1–8.
Nonaka I., *A Dynamic Theory of Organizational Knowledge Creation*, „Organization Science," 1994, t.5, nr 1, pp. 14–37.
Pai D.R., McFall K.S., Subramanian G.H., *Software effort estimation using a neural network ensemble*, „Journal of Computer Information Systems," 2013, t.53, pp. 49–58.
Paliwal M., Kumar U., *Neural networks and statistical techniques: A review of applications*, „Expert Systems with Applications", 2009, t.36, pp. 2–17.
Peeters P., Asperen J. van, Jacobs M., Vonk H., Others A., *The application of Function Point Analysis (FPA) in the early phases of the application life cycle A Practical Manual: Theory and case study*, NESMA 2005.
Perechuda K., *Zarządzanie wiedzą w przedsiębiorstwie*, Wydawnictwo Naukowe PWN 2005.
Piatetsky-Shapiro G., Frawley W.J., *Knowledge Discovery in Databases*, *Library Trends*, 1991, t. 48.

Pospieszny P., Czarnacka-Chrobot B., Kobyliński A., *Application of Function Points and Data Mining Techniques for Software Estimation - A Combined Approach*, „25th International Workshop on Software Measurement and 10th International Conference on Software Process and Product Measurement", Springer 2015, pp. 96–113.
Resolution Project, *CHAOS Summary 2009*, „Chaos," 2009, pp. 1–4.
Ruan D., Chen G., Kerre E.E., *Intelligent data mining: techniques and applications*, Springer Science & Business Media 2005, 5. wyd.
Ruchika Malhotra A.J., *Software Effort Prediction using Statistical and Machine Learning Methods*, „International Journal of Advanced Computer Science and Applications (IJACSA)," 2011, t. 2.
Schapire R.E., *The strength of weak learnability*, „Machine Learning," 1990, t.5, nr 2, pp. 197–227.
Schwalbe K., *Information Technology Project Management*, *Technology*, Course Technology, Boston 2014, t. 1.
Selby R.W., *Software engineering: The legacy of Barry W. Boehm*, „Proceedings - International Conference on Software Engineering," 2007, pp. 37–38.
Shukla R., Misra A.K., *Estimating software maintenance effort a neural network approach*, „Proceedings of the 2008 1st India Software Engineering Conference, ISEC'08," 2008, pp. 107–112.
Sobczyk M., *Statystyka*, PWN, Warszawa 2000.
Spalek S.J., *Critical Success Factors in Project Management -- To Fail or Not To Fail, That is the Question!*, „PMI Global Congress Proceedings," 2005, pp. 1–7.
Standish Group, *The CHAOS Manifesto 2011*, „The Standish Group International. EUA," 2011, p. 25.
Szupiluk R., *Dekompozycje wielowymiarowe w agregacji predykcyjnych modeli Data Mining*, Oficyna wydawnicza SGH, Warszawa 2013.
Taylor J., *Wstęp do Analizy Błędu Pomiarowego*, Wydawnictwo Naukowe PWN, Warszawa 1995.
Trendowicz A., *Software Project Effort Estimation: Foundations and Best Practice Guidelines for Success*, 2014.
Trocki M., Grucza B., Ogonek K., *Zarządzanie projektami*, Polskie Wydawnictwo Ekonomiczne 2009.
Villanueva-Balsera J., Ortega-Fernandez F., Rodríguez-Montequín V., Concepción-Suárez R., *Effort estimation in information systems projects using data mining techniques*, „Proceedings of the 13th WSEAS International Conference on Computers - Held as part of the 13th WSEAS CSCC Multiconference," 2009, pp. 652–657.
Weiß C., Premraj R., Zimmermann T., Zeller A., *How long will it take to fix this bug?*, „Proceedings - ICSE 2007 Workshops: Fourth International Workshop on Mining Software Repositories, MSR 2007," 2007.
Wen J., Li S., Lin Z., Hu Y., Huang C., *Systematic literature review of machine learning based software development effort estimation models*, „Information and Software Technology," 2012, t.54, pp. 41–59.
Xu L., Krzyzak A., Suen C.Y., *Methods of combining multiple classifiers and their applications to handwriting recognition*, „IEEE Transactions on Systems, Man, and Cybernetics," 1992, t.22, nr 3, pp. 418–435.
Zhou Z.-H., *Ensemble Methods: Foundations and Algorithms*, CRC Press, Boca Raton 2012.