Mining Competences of Expert Estimators

Hrvoje Karna *hkarna@fesb.hr*
Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture
University of Split, Split, Croatia

Sven Gotovac *gotovac@fesb.hr*
Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture
University of Split, Split, Croatia

Abstract

This paper reports on a study conducted with intention to identify competences of employees engaged on software development projects that are responsible for reliable effort estimation. Execution of assigned project tasks engages different human characteristics and effort estimation is integral part of development process. Competences are defined as knowledge, skills and abilities required to perform job assignments. As input data we used company internal classification and collection of employee competences together with data sets of task effort estimates from ten projects executed in a department of the company specialized for development of IT solutions in telecom domain. Techniques used for modeling are proven data mining methods, the neural network and decision tree algorithms. Results provided mapping of competences to effort estimates and represent valuable knowledge discovery that can be used in practice for selection and evaluation of expert effort estimators.
Keywords: Effort Estimation, Competences, Data Mining, Neural Networks, Decision Trees.

1. Introduction

Knowledge about the skills and competences of employees is of supreme importance for company success [22]. It enables efficient employee selection and staffing during project initialization as well as support during offer preparation and project planning [4]. Efficient competence management across organization ensures competitive position and a way to increase workforce productivity. Likewise structured competence framework provides transparency over available competences within organization and targeted development of those that are important through trainings and certifications [7].

Effort estimation is important part of software project management. The reliable effort estimates ensure planned project execution and compliance with the set time and budget constraints. Despite the long term efforts to produce accurate estimates based on formal and analogy based estimation methods expert estimation remains the most widely used technique of effort estimation [11]. Several reasons have contributed to this: studies consistently report that formal methods in comparison to expert estimation fail to produce more accurate estimates [9], expert estimation is easy to implement and finally expert estimation is more flexible regarding the type and format of the information used to produce estimates [19].

In this study we investigate the relationship between experts competences and accuracy of their effort estimates. To figure out the relation between ones competences and success in effort estimation we have to apply methods of knowledge discovery. Data mining algorithms are such an example and as studies report software engineering can benefit from use of this approach [15], [23]. Data mining in terms of software engineering consists of collecting software engineering data, extracting knowledge and when possible using this knowledge to improve the software engineering process. In this study we use two data mining approaches: neural networks and decision trees.

The remaining part of this paper is organized as follows: section 2 quotes the related research in this area. Section 3 introduces the model of competences used in the study. Section 4 describes the design of study. Section 5 explains the experiment setup and modeling performed in study. In section 6 survey results and their implications are discussed. Section 7 gives the conclusion and directions for the future research.

2. Related Research

Organizations have always been concerned about the competences of their employees. Today in a knowledge-based economy the success of organization mostly depends on workforce competences and competent employees are their main resource [2]. Competences are the best predictors of job performance [18]. In the same way estimating effort and therefore time and costs in different phases of a project is particularly important as these form a base on which decisions are made. The problem is when these estimates are not prepared by competent estimators. The present knowledge of how experts competences affect estimation accuracy arouses research interests [6].

Competence is a combination of knowledge, skills and process abilities that are causally linked and provide a base for job performance [16]. In certain form they represent a company's resource that could be exploited to gain competitive advantage [20]. While human resource development literature is mostly concerned with development of highly transferable generic competences that are required for most jobs or roles, particular company management is often emphasizing competences that are unique and company specific.

There are different competence models, usually in a form of a hierarchical catalogue that describes those that are desirable for organization and particular role [14]. Models depend on approach used to classify competences and can be one-dimensional or multi-dimensional which today are de facto standard [16]. Organizations use specialized IT-based systems to support the strategic competence management process [8]. Our previous study confirmed employees experience and role on a project give a high level notion of one's ability to successfully perform effort estimation tasks [12]. When it comes to competences required to perform estimation tasks questions are still incompletely answered.

This study was conducted with the aim to identify competences of professional software engineers engaged on projects within the company and occupying different positions that are important in determining one's ability to produce accurate estimates of efforts required to perform certain project tasks. Insights gained through application of various advanced knowledge discovery techniques help software engineers improve their everyday work practice.

3. Competences

Competency models are used to align individual capabilities with the competence of organization. These models are viewed as descriptive tools to identify the skills, knowledge, personal characteristics and behaviors that are required to efficiently perform a job in the organization [17]. The relation between competences and performance is shown on Figure 1.

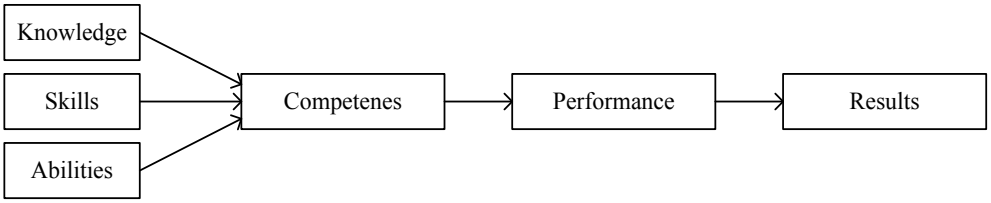

Fig. 1. Relation between competences and professional results.

A company competence model establishes a common language which allows better communication between project managers and employees as it defines job expectations. It can

also assist recruiting process where it can be used as some form of a guideline [13]. Knowing the skills, knowledge and abilities of employees allows better mapping of personnel to the company functions. For practical purposes of our study we are concerned with competences that a person working in a given occupational area should be able to do and demonstrate. Model of competences used by the company where study was performed covers tree segments: technical, professional and products and solutions competences. Each segment is further partitioned into sub-segments as it is shown in Table 1.

Table 1. Model of competences used in the study.

Segment	Competence	Description
Technical Competences	Operating Systems	Competence in use of operating systems
	Programming Languages	Competence in use of programming languages
	Development Environments	Competence in use of integrated development environments
	Database Systems	Competence in use of database management systems
	ALM Tools	Competence in use of application lifecycle management tools
	Project Process	Competence in application of different organization processes
Professional Competences	Development	Competence in different phases of software development process
	Operation and Maintenance	Competence in different operation and maintenance roles
	Project Types	Competence in various type of projects, current and past
	Role and Responsibility	Competence in relevant roles and responsibilities on projects, current and past
	Certifications	Level of certifications
Products and Solutions Competences	In-house Products and Solutions	Competence in development and use of in-house products and solutions
	Third Party Products and Solutions	Competence in development and use of third party products and solutions

It is important to note that in structured competence questionnaire used by the company to collect and store data each sub-segment represents an area that is further divided. For instance Programming Languages area specifically quotes languages in which skills are expected (C, C++, C#, Java, etc.) or Project Types area quotes current and past types of projects that employee possibly participated in (Maintenance, R&D, Product development, etc.).

The process of creating of competencies collection is organized the following way: initially the structured competencies questionnaire is created and distributed to all employees in the department. All employees have to fill the questionnaire and return it to responsible person. The method of estimation is therefore a self-assessment and competency in each specific area can be marked with levels noted in Table 2. Once all questioners are collected they are imported to central department competence database.

Table 2. Competence levels.

Level	Description
1 Initial	Performs routine tasks with supervision and guidance
2 Basic	Performs range of tasks, supervision is required for more complex tasks
3 Intermediate	Performs some complex and non-routine tasks, able to manage the subject without constant guidance, can oversee the work of others
4 Advanced	Performs a wide range of complex and non-routine tasks, can train others in this subject
5 Expert	Performs all tasks, applies a significant range of fundamental principles and techniques, has strategic view and can train others in this subject

4. Study Design

As it is mentioned the study was conducted in the Croatian branch of international company specialized for development of IT solutions used by a number of different telecom companies. This department has more than 50 employees occupying different positions of whom majority are software engineers responsible for software development and maintenance tasks on different projects. The solutions are developed using Microsoft technology stack (Team Foundation Server, Visual Studio, SQL Server, C#, etc.). In total 32 experts from 10 projects participated in study forming a set of 2090 items used for training and testing purposes. Details of projects included in study are displayed in Table 3.

Table 3. Details of projects included in study.

Project	Duration (months)	Development method	Team size	LOC[1]	Size[2]	Precedentedness[3]
1	20,40	Sequential	6	92.091	Small	True
2	26,66	Sequential	6	123.693	Small	True
3	34,15	Sequential	9	46.668	Small	False
4	31,90	Sequential	9	249.732	Medium	True
5	61,02	Sequential	12	457.745	Large	False
6	7,80	Sequential	12	167.644	Medium	True
7	27,11	Sequential	8	148.409	Small	True
8	17,01	Iterative	23	261.781	Large	False
9	34,94	Sequential	6	263.485	Large	True
10	66,37	Sequential	6	125.967	Small	True

[1] Size expressed in number of physical Lines of Code, calculated using LocMetrics tool (www.locmetrics.com)

[2] Company internal classification of project size (determined by financial indicators)

[3] Parameter that indicated presence of similar projects already executed in department

The work is organized in teams consisting of a project manager, software developers and testers. Solution architects, quality and configuration managers are department functions and engage in projects at different phases. From selected projects profile competences of in total 32 employees were randomly selected for later analysis. Characteristics of this competence data set are the following: a) out of 32 profiles 29 were males and 3 were females, b) roles occupied by employees in data set are: 4 project managers, 3 solution architects, 18 developers, 3 testers, 3 quality managers and 1 configuration manager and c) regarding the position level there were 16 seniors, 14 advanced and 2 junior engineers.

4.1. Data Sources

From the above listed projects development task and employee competence data required for the research were collected using following sources:

- Application lifecycle management tool implemented on projects that support development process. In this case it primarily served as a central place for collection of work item data. For this purpose on all considered projects Microsoft Team Foundation Server was used. Advantage that this and similar tools offer is the capability of various forms of data presentation, manipulation and export.
- The estimators competence data were gathered during company internal assessment procedure performed by dedicated department functions. The data collection was organized in form of a structured questionnaire that each employee received, had to fill and return to department. The questionnaire covered different aspects of employee profile of which major part was concerned with professional competences that are required to perform every day engineering tasks.

For employees involved on projects, collected competence data were structured in appropriate form, this made the total of 32 estimator profiles that entered the analysis. Input variables that are used to represent estimators competence characteristics are logically

organized into segments as defined in Table 1. Data exported from tracking system contain both reference to an item owner (employee) and assigned efforts. This allowed two things: first, linking of an item to estimators competence profile and second, calculation of estimation error. As a measure of estimation error i.e. accuracy the magnitude of relative error is used, MRE [3]:

$$MRE = \frac{abs(actual\ effort - estimated\ effort)}{actual\ effort} \quad (1)$$

The MRE (1) is the most widely used measure of effort estimation accuracy [1], [5], [21], it is basically a degree of estimation error in an individual estimate.

4.2. Data Mining Approach

Building of the data mining model considered in this research required the definition of business objectives. In this case it is the identification of the expert estimators competences and their relative importance in producing reliable effort estimates. This business objective was mapped to data mining objective with intention to create such a model that could later be implemented in practice. Methodological framework consists of following phases:

- Data collection: during which both work item and employee competence data were collected. This stage therefore included export of project tasks, identification of involved team members and structuring of their competence data.
- Data preparation: at this stage data was processed according to specific needs of model building process. The end product is data set that contains efforts data of each item and related employee (item was assigned to). This way single resulting data set from all ten analysed projects was generated. At this stage outliers, extremes and missing data are handled.
- Data partitioning: input data is randomly divided into two segments, training and test sets. From the initial data set the ratio of 2/3 of the data is used for the training (building of a model) and 1/3 for the testing phase (assessing of model performance).
- Model building: during this phase the predictive models are built using a MLP neural network and C&R decision tree algorithms and evaluated for predictive performance.

5. Experiment Setup

In accordance with the data mining practice data was prepared to produce input set comprising the total of 2090 records corresponding to projects being analyzed. Variables considered in the input data sets are listed in Table 4:

Table 4. Predictors and target in input data set.

Segment	**Variables**	
	Name (Code)	*Type*
Technical Competences	Operating Systems (OPS), Programming Languages (PRO), Development Environments (IDE), Database Systems (DBM), ALM Tools (ALM), Project Process (MET)	Predictor
Professional Competences	Development (DEV), Operation and Maintenance (OPR), Project Types (TYP), Role and Responsibility (ROL), Certifications (CER)	
Products and Solutions Competences	In-house Products and Solutions (IPS), Third Party Products and Solutions (TPS)	
	Magnitude of Relative Error (MRE)	Target

From the input set of variables 13 are used as predictors and single variable (MRE) as a target. Experiment was conducted using IBM SPSS Modeler 14.2. For analyzed data a stream representing data flow was developed to perform experiment. The experiments followed the

sequence in which data is initially fed into the stream after which it passed steps of preparation, transformation and partitioning before it entered the modeling element. The modeling elements used in this study implement following data mining algorithms:

- Neural network model uses MPL (multilayer perceptron) with the back propagation. Perceptron's architecture is organized into layers: input layer that receives information, hidden layer(s) and the output layer. During formation the model determines how the network connects the predictors to the target. This is done by hidden layer(s) that uses input values and modifies them using some weight. The activation function defines the output signal from the neuron. New value is then sent to the output layer where it is modified by some weight from connection between hidden and output layer. The back-propagation looks for the minimum of the error function. The combination of weights which minimize the error function is considered to be a solution of the learning problem.
- Decision tree model uses C&R (classification and regression) algorithm. Decision tree algorithm performs the procedure of examining the fields in dataset to find the ones that give the best classification or prediction by splitting data into subgroups. The process is applied recursively, splitting subgroups into smaller and smaller units until the tree is formed. The C&R algorithm minimizes the impurity at each step, where the node in the tree is considered "pure" if 100% cases in the node fall into a specific category of the target field. The output from a decision trees is a tree like structure that can be easily interpreted as a set of IF-THEN rules.

Application of data mining methods is well suited for our problem for several reasons. First of all they can operate on large data sets that are typical for research in field of software engineering. Next, they are used to extract knowledge from data and represent it in a form of rules for separation i.e. classification of input variable sets. This enables us to interpret and understand results of modeling. Finally, results from data mining process afterwards can be implemented in daily practice on projects, which can be a beneficial for business in multiple ways. In terms of our study these findings can enhance effort estimation process and thus result in more optimal utilization of project resources.

6. Survey Results

The outputs resulting from the models report the relative importance of the top predictors. The importance of each predictor is relative to the model and it identifies the input variables that matter the most during prediction process. Results of modeling process for both neural network and decision tree are displayed on Figure 2.

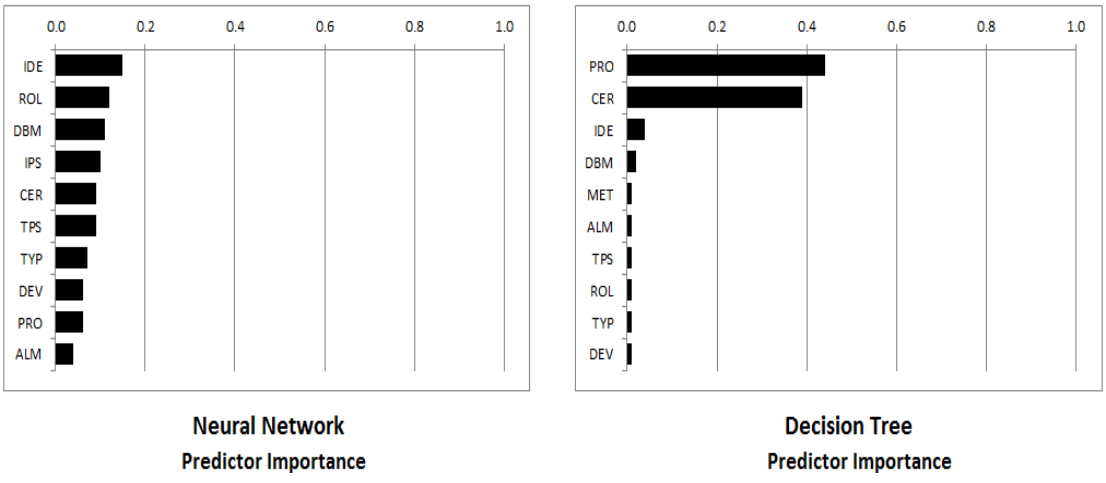

Fig. 2. Relative importance of predictors used in models.

The Multilayer Perceptron (MLP) neural network model returns the group of predictors with descending predictive power: IDE=0,15; ROL=0,12; DBM=0,11; IPS=0,10; CER=0,09; TPS=0,09; TYP=0,07; DEV=0,06; PRO=0,06; ALM=0,04. Resulting model has a single hidden layer with 10 neurons. Overall accuracy of resulting model is 57,9%. Although top predictors of estimation accuracy are competences i.e. know-how and skills in segments of development environment used on a project, current and previous roles and responsibilities, database management systems, in-house products and solutions know-how, certifications etc. from this model it is hard to designate typical predictors that could be used as classifiers.

On the other hand resulting model form the C&R decision tree clearly indicates predictors credible for the accurate effort estimates. This is obvious from distinctive values of their predictive importance: PRO=0,44; CER=0,39; IDE=0,04; DBM=0,02; MET=ALM=TPS=ROL=TYP=DEV=0,01. Model accuracy is similar to that of neural network. The resulting decision tree has depth = 3 and can be expressed as:

```
CER in [ "Basic" ]
        DBM in [ "Advanced" "Basic" ]
        DBM in [ "Intermediate" ]
CER in [ "Advanced" "Expert" "Initial" "Intermediate" ]
        PRO in [ "Advanced" "Basic" "Expert" ]
        PRO in [ "Intermediate" ]
                IDE in [ "Advanced" ]
                IDE in [ "Intermediate" ]
```

The decision tree can be interpreted the following way: the most important predictor of one's effort estimation accuracy is the competence CER. This competence belongs to professional segment and indicates the level of employees certification in areas important for assigned job position. In resulting model this predictor is rated with second greatest predictive importance. CER divides the initial set into two subsets, those with basic level of certification and the rest that belong to group with levels initial, intermediate, advanced and expert. First subset is further divided by DBM criteria based on its corresponding levels. The important segment of second subset is further divided by PRO, competence that indicates experts level of competence in programming languages, this is predictor with greatest importance in resulting model. Those with PRO level intermediate are later divided into subsets by IDE. To conclude, the decision tree gives us simple and readable form of results.

Results of modeling indicate competences that can be used as predictors of experts effort estimation accuracy. In terms of neural network model they are relatively closely grouped by predictor importance what made it hard to derive conclusions. On the other hand decision tree model gives comprehensive model from which a set of rules can be derived. Those rules, in terms of prediction of expert estimators accuracy can be expressed the following way: use level of certifications (CER) in areas relevant for the project context together with programming languages (PRO) competences as most relevant predictors. Only then consider group of predictors that is formed of IDE, DBM, IPS competences. Other predictors can be ignored due to their predictive power.

7. Conclusions and Future Directions

This paper reports a detailed description of the methodology used to develop predictive models in software engineering filed of effort estimation. Motivation comes from the need of introducing modeled approach of assessing expert competences used in effort estimation. The methodology was applied on the real data extracted from the tracking system used on projects and data collected in structured competence questionnaire. The study identified predictors that can be used to assess reliability of experts efforts estimates.

Results of this and future studies support the development of a model for enhanced expert effort estimation. Based on better understanding of effects that estimators competences have on reliability of effort estimates it would allow the application of corrective measures at early

stage of estimation process. Such a model is intended to enhance reliability of effort estimates and could be applied to everyday practice of software engineers.

References

1. Basha, S., Ponnurangam, D.: Analysis of Empirical Software Effort Estimation Models. International Journal of Computer Science and Information Security, 7 (3), (2010)
2. Chouhan V. S., Srivastava, S.: Understanding Competencies and Competency Modeling – A Literature Survey. International Organization of Scientific Research Journal of Business and Management, 16 (1), 14-22 (2014)
3. Conte, S. D., Dunsmore, H. E., Shen, V. Y.: Software Engineering Metrics and Models. Benjamin-Cummings, Menlo Park, CA (1986)
4. Dorn, J., Pichlmair, M., Schimper, K., Tellioglu, H.: Supporting Competence Management in Software Projects. Proceedings of International Conference on Concurrent Enterprising, 451-458 (2008)
5. Ferrucci, F. et al.: Genetic Programming for Effort Estimation an Analysis of the Impact of Different Fitness Functions. 2nd International Symposium on Search Based Software Engineering, Benevento, Italy, 89-98 (2010)
6. Giammalvo, P.: Creating a Competency Assessment for Cost Estimators. Project Management World Journal, 1 (2), (2012)
7. Hale, J.: Performance-Based Certification: How to Design a Valid, Defensible, Cost-Effective Program. Pfeiffer (2012)
8. Hustad, E., Munkvold, B., Moll, B.: Using IT for Strategic Competence Management - Potential Benefts and Chalenges. Proceedings of European Conference on Information Systems, (2004)
9. Jørgensen, M.: A review of studies on expert estimation of software development effort. The Journal of Systems and Software, 70 (1-2), 37-60 (2004)
10. Jørgensen, M., Sjøberg, D. I. K.: Impact of experience on maintenance skills. The Journal of Software Maintenance and Evolution, 14, 1-24 (2002)
11. Jørgensen, M., Boehm, B. and Rifkin, S.: Software Development Effort Estimation: Formal Models or Expert Judgment?. IEEE Software, 26 (2), 14-19 (2009)
12. Karna, H., Gotovac, S.: Estimators Characteristics and Effort Estimation of Software Projects. 9th International Conference of Software Engineering and Applications, Vienna, Austria, (2014)
13. Kenexa Corporation an IBM Company: How Competencies Enhance Performance Management. Available at: http://www.kenexa.com, (2012)
14. Klendauer, R. et al.: Towards a competency model for requirements analysts. Information Systems Journal, 22 (6), 475-503 (2012)
15. Layman, L. et al.: Mining Software Effort Data: Preliminary Analysis of Visual Studio Team System Data. Proceedings of the 2008 International Working Conference on Mining Software Repositories, 43-46 (2008)
16. Le Deist, D. F., Winterton J.: What is competence?. Human Resource Development International, 8 (1), 27-46 (2005)
17. Lucia, A.D.: Lepsinger, R., The Art and Science of Competency Models: Pinpointing Critical Success Factors in an Organization. Pfeiffer, San Francisco, CA (1999)
18. McClelland D.C.: Testing for Competence Rather Than for Intelligence. American Psychologist, 28 (1), 1-14 (1973)
19. Molokken K., Jørgensen, M.: Expert Estimation of Web-Development Projects - Are Software Professionals in Technical Roles More Optimistic Than Those in Non-Technical Roles, Empirical Software Engineering, 10, 7-29 (2005)
20. Nader, D. A., Tushman, M.: The Organization of the future: strategic imperatives and core competencies for the 21st century. Organizational Dynamics, 27 (1), 45-58 (1999)

21. Stensrud, E. et al.: A Further Empirical Investigation of the Relationship Between MRE and Project Size. Empirical Software Engineering, 8 (2), 139-161 (2003)
22. Winterton, J., Le Deist, F., Stringfellow, E.: Typology of knowledge, skills and competences: Clarification of the concept and prototype. Luxembourg (2006)
23. Xie, T., Thummalapenta, S., Lo, D., Chao Liu, C.: Data Mining for Software Engineering. IEEE Computer, 42 (8), 35-42 (2009)