J. Appl. Environ. Biol. Sci., 4(7S)479-489, 2014

© 2014, TextRoad Publication

ISSN: 2090-4274
Journal of Applied Environmental
and Biological Sciences
www.textroad.com

Feature Selection for Agile Development through Data Mining Techniques: An Application

Waqas Jawaid[1], Dr. Tahseen Ahmed Jilani[2], Yasar Methmood[3], Shah Muhammad[4]

[1,3,4] Department of Computer Science, Virtual University of Pakistan
[2]Department of Computer Science, University of Karachi

Received: September 1, 2014
Accepted: November 13, 2014

ABSTRACT

Traditionally software development has been performed by following a phased model. This model is called the waterfall (or traditional) model, in which the software development life cycle is divided into distinct phases i.e. requirements elicitation, software development, testing, and maintenance, which are followed in a defined order. The waterfall model creates heavy documentation, and most often results in huge rework and cost overruns because customer does not get the visibility of the project until it is very late. As an alternative to the documentation driven, heavyweight software development processes, many lightweight methodologies were created by software practitioners, e.g. crystal methodologies, feature driven development (FDD), scrum, extreme programming (XP) etc. The lightweight methodologies are called agile methodologies, and the development that is done following this model is called agile development. The agile methodologies follow the practices that add value to customers, and accept changes at any time during the development. From its inception, the agile development has been observed to be highly successful and it is used in most of the software companies now. However if a company wants to start using the agile development, then it is very difficult for that to choose which agile practices or features it should follow, because there are many practices for agile development, from which the company must choose which ones it would use. In this work the most important success factors will be extracted from the agile development practices that are successful in the software industry (with reference to the software practitioners in Pakistan). Through literature survey, the candidate factors will be selected, and then a survey will be conducted (using online survey forms) from the software practitioners of Pakistan, to find out which agile practices are most commonly used and which ones are the most successful for what kind of projects? The survey results will be analyzed using state of the art data mining techniques (e.g. dimension reduction techniques, clustering and classification techniques, and regression modelling) on multiple dimensions (extent of usage, extent of benefits) to find out which practices are most common and most successful among the software practitioners? The most common and useful success factors/features of agile development will be extracted on the basis of the obtained results.

KEYWORDS: Agile Development, Data Mining techniques, fuzzy-c mean clustering, Extreme Programming, Scrum, Feature Driven Development

1 INTRODUCTION

Usually software development has been performed by following a phased model. This model is called the waterfall (or traditional) model, in which the software development life cycle is divided into distinct phases i.e. requirements elicitation, software development, testing, and maintenance, which are followed in a defined order. The waterfall model creates heavy documentation, and most often results in huge rework and cost overruns because customer does not get the visibility of the project until it is very late. As an alternative to the documentation driven, heavyweight software development processes, many lightweight methodologies were created by software practitioners, e.g. crystal methodologies, feature driven

* **Corresponding Author:** Waqas Jawaid, Department of Computer Science, Virtual University of Pakistan

development (FDD), scrum, extreme programming (XP) etc. The lightweight methodologies are called agile methodologies, and the development that is done following this model is called agile development.
Agile development has been observed to be very successful from its beginning. However there are many different agile development methodologies and practices which are being used by software development companies. If a company wants to start using agile development then it must choose which agile practices it should follow from the many available practices. We know that there are some key factors in any work that are vital for its success. These factors are known as critical success factors. Hence, we should attempt to find out which practices are the key success factors or features of agile development, so that a company that wants to start using agile development, can focus on these features which are vital for the success of agile development, rather than wasting its time and resources on trying out the less useful practices. Since there are a large number of agile practices, so we would need to use the data mining techniques to dig out the useful information from the vast available data. Feature Selection refers to the problem of identifying a subset of features from a feature set that can be used to create a classification model for a certain task. In this work, the feature selection for agile development will be done using the data mining techniques. To our knowledge, no prior work of this kind has been done on agile development practices being used in Pakistan.

1.1Introduction to Agile Development

The term Agile development was formally defined in 2001 in the Agile Manifesto, when about seventeen renowned software process practitioners met in ski resort North America and agreed on a set of principles which are at the core of agile software engineering [1]. Even before the agile manifesto, the software practitioners were practicing the practices that were mentioned in the agile manifesto, but these were not well defined. Those practitioners were working separately to create different lightweight software development techniques but finally they unified their thoughts to come up with a common set of principles to define agile development, which is known as the Agile Manifesto [1]. At that time the usual software development was very much focused on control mechanisms and planning. Software development was not very efficient at that time. It was very time consuming, and most often it produced large documentations and other artifacts which were never used afterwards. Agile software development aims at reducing the wasted effort and resources, and focuses on improved communication and business value for customers and quick response to any changes requested by customers. Many software development techniques are classified as agile development approaches. The agile development creates software processes that quickly respond to requirements changes and depend on communication and collaboration (more than the documentation) for requirements gathering and other processes. The supporters of agile development techniques claim that the agile development is the response of software practitioners to the flaws of traditional software development techniques, but this claim has not been researched thoroughly. More research is needed to determine various aspects of agile development and its ease of use or negative impacts on software development processes. The impact of various agile development practices also needs to be researched for various software domains and its applicability according to organization size.

1.2 Drawbacks of Traditional Development Methodologies

The basic difference in agile and traditional methodologies is that the agile methodologies accept that the changes in software are inevitable and even vital for the project's success. It is critical to react to changes in a timely manner to make a project successful. In traditional methodologies, the requirements are fixed before the development is started, and changes are discouraged or denied. The main point in agile methodologies is that the changes can be made at any stage of development. The founders of the agile manifesto, Jim Highsmith and Martin Fowler stated that it is more effective to facilitate change than trying to stop it [2]. Jones and Boehm also mentioned that in their careers, they observed that the requirements change at the rate of 25% or even greater [2][3].

Standish group [4] conducted a research with 365 respondents, and collected data about 8,380 software projects from the industry. They observed that only 16% of the projects can be said to be successful, with on-budget and on time delivery and all the specified features included. About 53% projects were over-time and over-estimated budget and had less features than originally anticipated while 31 projects were completely cancelled before being completed. It was also noted during the study that the major causes for a project's success are clearly stating the requirements, support from executive management and user involvement with the development team [4].

The Standish group also made another important observation that almost 45% features of a software are not used by the users. This encourages the developers to make the code and design as simple and small as possible, so that only those features are added which are needed by the customers, because most of the coding and complexity is not in fact required by the users.

2 Introduction to Data Mining Techniques

The technique to find meaningful and useful information (also called nuggets) in databases is called data mining. There are various algorithms and methods for data mining. Data mining is usually divided in two categories, descriptive and predictive. The descriptive model is used to identify patterns in some data, and the predictive model is used to make predictions about the new results through the analysis of obtained results from previously collected data [5].

The data mining techniques described below were applied on the features that were obtained through our survey.

2.1 Principal Components Analysis

Principal Components Analysis (PCA) is a technique that can be used for dimensionality reduction, i.e. if the data has many dimensions and we need to reduce these then this technique can be used. It is also known as Karhunen-Loeve method (or K-L method). This is one of the most valuable and popular statistical techniques. Let us assume that the data set that we have, has many data vectors consisting of n dimensions or attributes, then the PCA creates *m*-dimensional orthogonal vectors that show the data in a way that retains the maximum information but reduces the dimensions, because m ≤ n. Hence the data is then reduced to less dimensions than the original data, but still retains the most significant information.

The original data is converted in new dimensions through PCA. The new variables are represented as a linear combination of the old variables, which is shown as follows [5]:

$$Z_1 = b'_1Y = b_{11}Y_1 + b_{12}Y_2 + \cdots + b_{1m}Y_m$$

$$Z_2 = b'_2Y = b_{21}Y_1 + b_{22}Y_2 + \cdots + b_{2m}Y_m$$

.....

$$Z_p = b'_pY = b_{p1}Y_1 + b_{p2}Y_2 + \cdots + b_{pm}Y_m$$

The data can be written as a matrix also, in matrix from, it can be written as Z=BY, where the b parameters are the loading parameters. The newly obtained axes are made orthogonal to each other by adjusting these, to maximize the information gain.

$$Var(Z_i) = b'_i \sum b_i \quad . \quad i = 1.2....p$$

$$Cov(Z_i.Z_K) = b'_i \sum b_K \quad . \; i = 1.2....p$$

The first principal component has the highest variance. For feature transformation, firstly the covariance matrix U is determined, because directly computing the matrix B is not possible. U can be shown as

$$U_{m\times n} = \frac{1}{m-1}\left[\sum_{i=1}^{m}(Y_i - \bar{Y})'.(Y_i - \bar{Y})\right].$$

$$\text{where } \bar{Y} = (\frac{1}{m})\sum_{i=1}^{m} X_i$$

After that the Eigen values are calculated for the covariance matrix U. Lastly, we define a linear transformation from m dimensions to n dimensions (where n<m) [5].

2.2 Regression Model

Through regression, future outcomes can be predicted based on old data. The power of relationship among 2 variables may be calculated by using the bivariate model. The linear regression model has the following general form.

$$z = a_0 + a_1 y_1 + \cdots + a_m y_m + \epsilon$$

Where y_1, y_2,… , y_m are known as regressors, which are the input variables. ϵ is a random number. a_0, a_1, …, a_m are constants that are selected on the basis of statistical methods according to the input sample values. This model is also called the multiple linear regression model, because there are many variables involved. This title represents that this model lies in hyper dimensional space. There are certain observed values that are not in accordance with the expected values, which are known as outliers. Most of the times, the data is preprocessed to analyze the outlying values and interferences [5, 7].

2.3 Logistic Regression Model

Logistic regression model is used to model the probability of an instance as a function of linear predictor variables. This model may be defined as follows:

$$E(Z/x) = \frac{e^T}{1+e^T} = \pi(x) = \frac{e^T}{1+e^T}$$

Here $\pi(x)$ means the expected outcome of the response variable, e is natural log base, while T can be represented as:

$$T = \rho_0 + \rho_1 X_1 + \rho_2 X_2 + \ldots + \rho_h X_h$$

Where ρj are coefficients and Xj are predictors for h predictors, while j=1, 2,…, h [7, 8].

2.4 Cluster Analysis

Clustering is the procedure of combining a group of abstract or physical objects in classes of analogous objects. The classes obtained are known as clusters. The objects in the same cluster can be considered as one set. Even though classification is also an efficient method to separate classes or groups of objects, it needs many large training patters which are used by the classifiers to create separate groups from the data. The large training patterns make the classification process more expensive than clustering. Hence it is usually considered more effective to create clusters from objects (clustering) rather than partitioning one large group.

Data clustering is being developed vigorously. There are various areas of research in clustering, including marketing, biology, machine learning, spatial database technology, statistics, and data mining. Clustering is a kind of unsupervised learning or learning by observation i.e. clustering does not depend on pre-created classes and class labeled training examples [6, 9].

2.5 Fuzzy C Mean Clustering

The Fuzzy C Mean clustering algorithm (FCM) is a very famous clustering technique. FCM is believed to be the most common fuzzy clustering technique. There are several improvements proposed for the FCM, e.g. FCM based on Cluster Density (FCM-CD) by [10]. The classic FCM considers the Euclidean distance to determine the similarity of data points that implies similar partition tendencies for data sets. Another limitation is that the performance of clustering algorithm is greatly influenced by the cluster density and shape. Lou *et al.* [10] proposed a distance regulatory factor to fix the similarity issue, which shows the distribution of cluster data points. This factor is used for distance correction after applying the traditional FCM [10, 11].

In FCM the objective is to get a fuzzy c-partition for a data set by minimizing an objective function J_{FCM},

$$J_{FCM}(\boldsymbol{X},\boldsymbol{U},\boldsymbol{V})| = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m} d_{ij}^{2}$$

as demonstrated below. (1)

The constraints on the fuzzy membership degree are as follows

$$s.t. \quad u_{ij} \in [0,1], \sum_{j=1}^{n} u_{ij} > 0, \sum_{i=1}^{c} u_{ij} = 1$$

The optimal value of J_{FCM} is obtained by Alternative Optimization method. The Alternative Optimization method depends upon the cluster prototypes and fuzzy membership degrees.

The fuzzy membership degrees are given by the following equation

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{\frac{2}{m-1}}}$$

The cluster prototypes are given by

$$\boldsymbol{v}_i = \frac{\sum_{j=1}^{n} u_{ij}^{m}\boldsymbol{x}_j}{\sum_{j=1}^{n} u_{ij}^{m}}$$

The distance between data points is calculated as

$$d_{ij}^{2} = \|\boldsymbol{x}_j - \boldsymbol{v}_i\|_A^2 = (\boldsymbol{x}_j - \boldsymbol{v}_i)\boldsymbol{A}(\boldsymbol{x}_j - \boldsymbol{v}_i)^T$$

In the above equation A is a positive definite matrix which is symmetric. In classic FCM, Euclidean distance is considered as the distance measure. In that case A=I in the above equation [10].

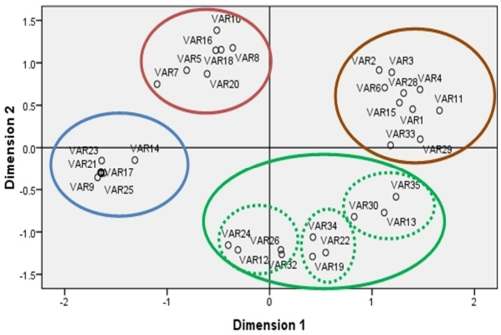

Figure 1. Clustering Example

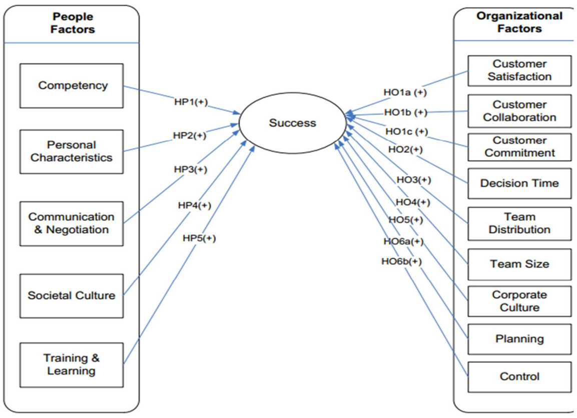

Figure 2. Factors that contribute to the success of agile project

The results obtained through the research are described below.

3.1Candidate Success Factors for Agile Development

There are numerous practices in agile development that are being used in software development companies. Since our aim is to find out the most important agile development features (or success factors), we must first identify the candidate success factors through literature review. Afterwards, we can obtain the success factors from the candidate factors, through a survey. Hence a thorough literature review was performed, during which many articles were explored to discover the agile development practices being used in industry. The agile development practices, which are considered as the potential success factors, were identified and a survey was conducted from the software practitioners in Pakistan, to identify the features for agile development that are vital for a project's success.

3.2Survey Conducted from Software Practitioners

The mentioned agile practices were extracted, and an online survey questionnaire was made, in which the candidate success factors and practices of agile development were mentioned. The respondents were asked to rate the mentioned agile development practices according to how rigorously they applied each of the mentioned practices in their successful agile projects, and how beneficial did they find those practices? The survey was conducted among the software practitioners (project managers, software engineers, and quality assurance engineers) from various software houses, ranging from freelancers to the companies having more than 500 employees. The response received was overwhelming and numerous completed responses were received.

3.3Principal Components Analysis

In our survey, the data was collected for about 61 agile development features/practices. To reduce the dimensionality of data, Principal Components Analysis (PCA) was applied. Using PCA, the features were mapped on 50 principal components. From the principal components, the features were found which had the most impact. In this way, the features that had the most control over the data were obtained. Using PCA, 61 features were reduced to 30 features. The reduced feature set is given below.

3.4 Multiple Linear Regression

After the application of PCA, the feature set was reduced to 30 features. Then the multiple linear regression was applied to observe the features which had strong impact on the overall success of an agile development project. For the regression model, the development features were considered as independent variables in the model, whereas the overall success of the project was taken as the dependent variable. The results obtained through the multiple linear regression are shown in the table 2 and table 3.

In the multiple regression model, our criteria for important features is such that the sig. value should be less than or equal to 5%, and the t value should be greater than or Equal to 2. Among the features that fulfill the above criteria, the impact of a feature on the overall success (the dependent variable) is determined by the Beta value for that feature.

Table 1. AGILE development features/practices selection using PCA.

• **Planning game**	• **Customer involvement**	• **Incremental design**
• **Efficient and effective communication**	• Preferring working software over documentation	• Starting each release with planning and ending with a review
• **Pair programming**	• Continuous integration	• Decide as late as possible
• **Customer collaboration**	• Metaphors	• Coding standards
• **Simplicity of design**	• Empowering the project team	• Sit together
• **Use of tools**	• Incremental development	• User stories
• **Project schedule**	• Negotiated scope	• Societal culture
• **Control**	• Team continuity (same members continue to be on the team for a project)	• Reflective improvement (processes are improved continuously)
• **Retrospectives**	• Customer satisfaction	• Agile team culture
• **Personal characteristics**	• Single code base	• Daily meeting

The regression model showed that total 9 features have positive impact on the overall success of the project. Hence those features are the most important features (or success factors) for agile development projects. The most important features obtained by the regression model are listed below, in the order of importance. the importance of each variable below is computed via its beta value and t-test. The overall model testing is performed using forward selection regression modelling.

Table 2. ANOVA For Regression Model

	S.S.	D.F.	M.S.E	F-Ratio	P-value	R Square	Adjusted R Square
Regression	147.531	29	5.087	10.053	0.000	0.921	0.829
Residual	12.651	25	0.506				
Total	160.182	54					

Table 3. Features having the Most Impact on Project Success

Feature	Beta	t	Sig.
Daily meeting	1.3	8.442	0
Personal characteristics	0.85	4.418	0
Empowering the project team	0.506	4	0
Retrospectives	0.487	4.37	0
Efficient and effective communication	0.485	3.63	0.001
Pair programming	0.431	3.367	0.002
Customer collaboration	0.307	2.207	0.037
Agile team culture	0.303	2.812	0.009
Customer involvement	0.247	2.35	0.027

3.5FCM Clustering Analysis

The features that were obtained after applying PCA were used as input for FCM Clustering. Through FCM, the data was clustered in 10 clusters. For each cluster, the difference (distance squared) was calculated for each attribute with cluster centers. In this way, the features were obtained that had the least distance with the cluster centers. The features that have least distance with the cluster centers are considered as important features. The features that were obtained as significant through the FCM are shown below in table 4. The features that are found to be significant according to PCA and FCM are shown in table 5.

Table 4. Significant Features According to FCM

Rank (FCM)	Feature
1	Personal characteristics
2	Societal culture
3	User Stories
4	Incremental design
5	Continuous integration
6	Planning game
7	Project schedule
8	Negotiated scope
9	Empowering the project team
10	Simplicity of design

Table 5. Significant Features According to FCM and PCA

Serial #	Feature	FCM Rank	PCA Rank
1	Personal characteristics	1	1
2	Continuous integration	2	7
3	Planning game	3	6
4	Empowering the project team	9	2
5	Simplicity of design	10	8

4 Conclusion

We aimed to identify the agile development techniques that play the most vital role for the success of an agile development project, so that those agile development practices can be used by the organizations or practitioners who want to adapt to agile development. That work was needed because there is a lack of empirical evidence on the most successful agile development practices in the industry. To begin with, various agile development practices were identified by literature review of already published articles on agile development. Afterwards, a survey was conducted that targeted agile development practitioners, to obtain the agile development practices that have most widely been used in the software industry. We observed various agile development features or practices that have been practiced in the software development industry (especially in Pakistan). Then we explored various data mining techniques. We used several data mining techniques, including Principal Components Analysis (PCA), Multiple Regression Analysis, Cluster Analysis and FCM Clustering. We applied those data mining techniques on the survey results, to obtain the most important agile development features/practices. The selected agile development features are mentioned below.

1. Daily meeting
2. Personal characteristics
3. Empowering the project team
4. Retrospectives
5. Efficient and effective communication
6. Pair programming
7. Customer collaboration
8. Agile team culture
9. Customer involvement

In addition to the above practices, the following practices were also found to be important for the success of an agile development project.

1. Continuous integration
2. Planning game
3. Simplicity of design

REFERENCES

1. Hansson. C, Dittrich. Y, Gustafsson. B, and Zarnak. S, “How Agile are Industrial Software Development Practices”, The Journal of Systems and Software 2006, 79(9): 1295-1311.
2. Boehm. B. W, and Philip. P, “Understanding and Controlling Software Costs”, IEEE Transactions on Software Engineering 1988, 14(10): 1462-1477.
3. Jones. C, “Applied Software Measurement: Assuring Productivity and Quality”. 1st Ed, McGraw-Hill, New York, USA, 1997.
4. Awad. M. A, “A Comparison between Agile and Traditional Software Development Methodologies”, Bachelors Thesis, School of Computer Science and Software Engineering, University of Western Australia, Australia, 2005.
5. Yasin. H, Jilani. T. A,and Danish. M,“Hepatitis-C Classification Using Data Mining Techniques”, International Journal of Computer Applications 2011, 24(3): 1–6.
6. Han. J, & Kamber. M,“*Data Mining: Concepts and Techniques”,*2nd Ed, Morgan Kaufmann Publishers, San Francisco, USA, pp. 39-40. ISBN: 1-55860-901-6. 2006.
7. Walpole. R. E, Myers. R. H, Myers. S. L, & Ye. K,“Probability and Statistics for Engineers”, 9th Ed. Prentice Hall, Upper Saddle River, New Jersey, USA, pp. 100-115. ISBN: 978-0321629111. 2011.
8. Jilani. T. A, Yasin. H, Yasin. M, & Ardil. C, “Acute Coronary Syndrome Prediction Using Data Mining Techniques - An Application”. International Journal of Computational Intelligence 2009, 5(4): 295-299.
9. Jilani. T. A, & Burney. S. M. A, “Multiclass Bilateral-Weighted Fuzzy Support Vector Machine to Evaluate Financial Strength Credit Rating”. International Conference on Computer Science and Information Technology 2008, (Vol.1, pp. 342-348).
10. Lou. X, Li. J, & Liu. H, “Improved Fuzzy C-means Clustering Algorithm Based on Cluster Density”. Journal of Computational Information Systems 2012, 8(2): 727-73
11. Jilani. T. A, & Naqvi. S. A. R, “A Review of Probabilistic Graph Models for Feature Selection with Applications in Economic and Financial Time Series Forecasting”, VFAST Transactions on Software Engineering 2014, 3(1): 7-14.