Uncertainty Assessment in Project Scheduling with Data Mining

C. Capa*, K. Kilic, G. Ulusoy

Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey

(c_capa@encs.concordia.ca, gunduz@sabanciuniv.edu, kkilic@sabanciuniv.edu)

Abstract–During project execution, especially in a multi-project environment project activities are subject to risks that may cause delays or interruptions in the baseline schedules. This paper considers the resource constrained multi-project scheduling problem with generalized activity precedence relations requiring multi-skilled resources in a stochastic and dynamic environment present in the R&D department of a home appliances company and introduces a two-phase model incorporating data mining and project scheduling techniques. This paper presents the details of Phase I, *uncertainty assessment phase*, where Phase II corresponds to *proactive project scheduling* module. In the proposed uncertainty assessment approach models are developed to classify the projects and their activities with respect to resource usage deviation levels. In doing so, the proposed approach enables the project managers not only to predict the deviation level of projects before they actually start, but also to take needed precautions by detecting the most risky projects. Moreover, Phase I generates one of the main inputs of Phase II to obtain robust baseline project schedules and identifies the risky activities that need close monitoring. Details of the proposed approach are illustrated using R&D project data of a leading home appliances company. The results support the efficiency of the proposed approach.

Keywords**:** decision analysis, project management, risk analysis, uncertainty modeling

*Currently at the *Faculty of Engineering and Computer Science, Concordia University, Montreal, Canada.*

May, 2015.

1. INTRODUCTION

A major mode of production of goods and services involves projects and hence, project management and scheduling. A large number of firms and organizational units are organized as project-based organizations such as technology firms,consulting firms and R&D departments among others perform almost all their work through projects. These organizations operate generally in a multi-project environment operating on more than one project simultaneously. These projects are interrelated since the same pool of resources is employed to execute them. Therefore, generating project schedules has become more of an issue to better utilize resources to achieve project objectives.Such a schedule helps to visualize the project and is a starting point for both internal and external planning and communication. Careful project scheduling has been shown to be an important factor to improve the success rate of projects.

Most of the studies in project scheduling literature assume complete information about the problem and develop scheduling methodologies for the static and deterministic project scheduling problem (see, e.g., Demeulemeester and Herroelen, 2002; Hartman and Briskorn, 2010). However, uncertainty is inherent in all project management environments. In reality, the situation is dynamic in the sense that new projects arrive continuously and stochastic in terms of inter-arrival times and work content. Furthermore, during project execution, especially in a multi-project environment project activities are subject to uncertainty that can take many different forms. Activity duration estimates may be off, resources may break down, work may be interrupted due to extreme weather conditions, new unanticipated activities may be identified, etc. All these types of uncertainties may result in a disrupted schedule, which leads in general to the deterioration of the performance measures. Thus, the need to protect a schedule from the adverse effects of possible disruptions emerges. This protection is necessary because a change in the starting times of activities could lead to infeasibilities at the organizational level or penalties in the form of higher subcontracting costs or material acquisition and inventory costs. Hence, being able to generate robust schedules becomes essential if one aims at dealing with uncertainty and avoiding unplanned disruptions.

Most widely used approach to handle uncertainty is to attach it into activity durations without explicitly considering the sources of risks. In this activity-based approach, uncertainty is attached to activity durations using three-point estimates of low, most likely and long activity durations and assuming appropriate probability distributions (Hulett, 2009). However, this approach fails to assess the impact of risks individually on the activity durations. Therefore, it is of interest to develop risk integrated project scheduling techniques to produce robust baseline schedules, i.e., baseline schedules that are capable of absorbing such disruptions. This makes risk-based uncertainty assessment an essential step for project scheduling.

This paper is motivated from a preemptive resource constrained multi-project scheduling problem (preemptive-RCMPSP) with generalized activity precedence relations requiring multi-skilled resources in a stochastic and dynamic environment present in the R&D department of a home appliances company. A two-phase model is developed incorporating data mining and project scheduling techniques to schedule these R&D projects. In this paper, our focus will be limited to the Phase I, *uncertainty assessment phase*, of the developed two-phase framework. Details of Phase II, *proactive project scheduling phase*, can be found in Capa and Ulusoy (2015). Proposed *uncertainty assessment phase*provides a systematic approach to assess risks that are thought to be the main factors of uncertainty and measuring the impacts of these factors to durations by utilizing the most important data mining techniques: *feature subset selection, clustering,*and *classification*.

Next section introduces briefly risk analysis in project scheduling and risk integrated project scheduling methods presented in literature. In Section 3, the problem on hand and the problem environment are explained in detail. The details of the *uncertainty assessment phase* of the developed two-phase methodology are presented in Section 4. In Section 5, a real case is introduced andutilized for presenting the details of the developed methodology. Finally in Section 6, the paper is concluded with a discussion on the results of the case study, final remarks and possible future research agenda.

2. RISK ANALYSIS AND RISK INTEGRATED METHODS IN PROJECT SCHEDULING

Risk is defined as an uncertain event or condition that, if it occurs, has a positive or negative effect on a project objective (PMI, 2000). The goal of risk analysis is to generate insights into the risk profile of a project and use these insights in order to drive arisk response process. In literature, risk analysis process is divided into four main sub-processes: *risk identification, risk prioritization, quantitative risk assessment* and *quantitative risk evaluation* (Herroelen et. al., 2014). A wide body of knowledge on quantitative techniques has been accumulated over the last two decades. *Monte Carlo Simulation* is the predominant technique so far both in practice and in literature. *Proactive project scheduling* has recently emerged as another approach of interest.

With risk information on hand, *proactive project scheduling* aims at constructing a *stable*initial baseline schedule that anticipates possible future disruptions by exploiting statistical knowledge of uncertainties that have been detected and analyzed in the project planning phase. *Stability* is a particular kind of robustness that attempts to guarantee an acceptable degree of insensitivity of the initial baseline schedule to the disruptions that may occur during the project execution and it represents the degree of the difference between the baseline and realized schedule. Although it can be represented by a number of ways such as the number of disrupted activities, or the number of times that an activity is re-planned, the most widely used measure is the *stability cost function* which is the expectation of the weighted sum of the absolute percent deviation (%deviation) between the planned and realized activity starting times. The activity dependent weights in this *stability cost function* represent the marginal cost of deviating the activity's starting time from the scheduled starting time and it reflects either the difficulty in shifting the booked time window for starting or the importance of on-time performance of the activity. They may include unforeseen storage costs, extra organizational costs, costs related to agreements with subcontractors or just a cost that expresses the dissatisfaction of employees with schedule changes (Van de Vonder et al. (2007)). The objective of the proactive project scheduling is then to minimize the expected absolute %deviation between the planned and realized activity start times. Since the analytic evaluation of this expected value is burdensome, a natural way out is to evaluate it through simulation, which mimics the project execution over a number of scenarios. For more details on stability in project scheduling we refer to Leus (2003) and Leus and Herroelen (2004).

Although there are some efforts to develop risk integrated project scheduling techniques to produce robust baseline schedules, the literature on this subject is very scarce. Jaafari (2001), Kirytopoulos et al. (2001), Schatteman et al. (2008), Creemers (2011), and Herroelen (2014) are notable examples of the risk integrated project scheduling methodologies. Jaafari (2001) presents an integrated and collaborative approach, which setsthe life cycle objective functions as the basis of evaluation throughout the project life cycle. Kirytopoulos et al. (2001) introduce a knowledge system to identify risks and their assessments in project schedules. Expert knowledge,checklists and risk breakdown structure are utilizedin the system. Schatteman et al. (2008) present a computer supported risk management system that allows identification, analysis, and quantification of the major risk factors and the derivation of their probability of occurrence and their impact on the duration of the project activities. Creemers et al. (2011) show that a *risk-driven* approach is more efficient than an *activity-based* approach when it comes to analyzing risks. In addition, the authors propose two ranking indices; one activity-based index that ranks activities and a risk-driven index that ranks risks. These indices allowidentifying the activities or risks that contribute most to the delay of a project to assist project managers in determining where to focus their risk mitigation efforts. Herroelen (2014) proposed a risk integrated tabu search methodology that relies on an iterative two-phase process. While in phase one, the number of regular renewable resources to be allocated to the project and the internal project due date are determined, phase two implements a proactive-reactive schedule generation methodology through time and/or resource buffers.

3. PROBLEM DEFINITION AND ENVIRONMENT

The problem on hand is scheduling of the research and development (R&D) projects with a priori assigned resources in a stochastic environment present in the R&D Department of a home appliances company. In the R&D Department, R&D projects related to technologies to be employed for the current and future product portfolio are conducted. The problem environment under consideration contains multiple projects consisting of activities using multi-skilled renewable resources. The project networks are of activity-on-node (AON) type with finish-to-start (FS) and start-to-start (SS) precedence relations with zero and positive time lags. No precedence relation is assumed between projects. All the projects are managed with a *stage-gate* approach.

The R&D Department is organized in technology departments and these technology departments are comprised of technology families, each of which works on a different technology field (*Fluid Dynamics, Material Science, Thermodynamics, Cleaning, Vibration and Acoustics, Structural Design, Power Electronics*, and *Electronic Assessment*). Each technology family has a technology family leader, who is responsible for all the researchers and technicians working in the corresponding technology family. Prior to the initiation of a project, the activities of the project as well as the precedence relations between activities are determined by the project leader and the project team. Since it is difficult to make correct estimations on the work content of the activities in the planning phase, they are defined as aggregate activities, which might include several subtasks that might be detailed in a later time. There are two types of resources in the R&D Department: *human resource* and *equipment*. Human resources consist of researchers and technicians. All the equipment, machines, mechanisms and laboratories are included under the equipment category. Human resources are multi-skilled, i.e., each human resource has its own specialty and the degree of that specialty differs from one human resource to another. This makes human resources critical for the R&D Department since the human resources are not necessarily substitutable. Each activity requires resources from different departments and different technology families for certain working hours. Thus, the projects are conducted in a multi-disciplinary environment. Moreover, resources that an activity requires do not need to work together or simultaneously. They can even stop working on that activity for a while and then continue later, i.e., the work of the resources on activities are pre-emptive leading to pre-emptive activities. The resource requirement of activities and hence, the durations of the activities are uncertain. In the literature, the activities require a number of resources for certain deterministic or stochastic durations instead of requiring working hours, which is the case here. Therefore the project environment considered in this paper is different than the project scheduling environments existing in the literature in the sense that it has different data requirement.

The problem on hand can be considered as an extension of the resource constrained multi-project scheduling problem (RCMPSP) with generalized precedence relations and multi-skilled resources to include pre-emption, stochastic activity duration and resource availabilities, and dynamic arrival of projects. The objective is, by considering the possible activity %deviations beforehand, generating stable baseline project schedules with an acceptable makespan.

4. PROPOSED METHODOLOGY FOR UNCERTAINTY ASSESSMENT

The main purpose of this study is topresent the *uncertainty assessment phase* of an integrated methodology for robust project scheduling. The uncertainty assessment phaseprovides a systematic approach to assess uncertainty by identifying the most important factors of uncertainty, measuring the impacts of these factors to the resource usage %deviation levels of projects and their activities and generating activity %deviation distributions. This phase is designated as Phase I of a two-phase model for robust project scheduling. In Phase II, *proactive project scheduling phase*, we use a bi-objective GA employing two different chromosome evaluation heuristics to generate robust baseline schedules. This bi-objective GA provides a set of robust non-dominated baseline schedules for scheduled activities to the decision maker. The decision maker can then choose one of these non-dominated robust baseline schedules to be used as the main baseline plan for the activities considering the dynamics of the current project management environment. This baseline plan is then used as a reference point in the implementation and monitoring phase of the projects and can be revised, if needed. Basic framework of the two-phase approach is given in Figure 1.

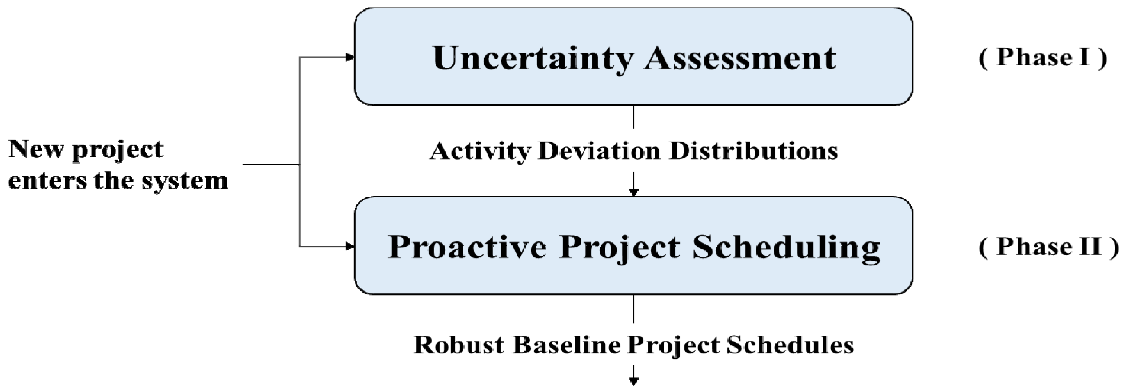

Figure 1. Framework of the two-phase approach

Risk and uncertainty assessment is an essential step in proactive scheduling. Without a risk analysis and an uncertainty assessment mechanism, failure in the management of projects is inevitable in competitive and stochastic multi-project management environments. Although risk factors and their probability of occurrence together with possible impact levels are estimated in the aforementioned R&D Department, defined risks are not associated with the project activities. Lack of this important component of risk data precludes the implementation of *risk-driven* approaches similar to those that are proposed in the literature for robust project scheduling (e.g., Herroelen et al., 2014). For that reason, in this study uncertainty is assessed with the help of data mining tools. The main source of uncertainty in activity durations is the uncertainty resulting from resource usage %deviations from estimated levels for the activities. Therefore, in the uncertainty assessment procedure presented in this paper, we investigate the %deviations of resource usages of projects and their activities.

Proposed uncertainty assessment phase suggests assessing the uncertainty of projects and their activities by first classifying the "projects" with respect to their percent resource usage %deviations, then classifying the "activities" with respect to their percent resource usage %deviations, thus constructing a resource usage %deviation assignment procedure for the prediction of %deviation levels of the activities of a newly arrived project. From now on, "percent resource usage %deviation" will be referred to as "%deviation". Final output of this phase is %deviation distributions for the activities of projects. It should be noted that, this phase is not problem-specific, i.e., can be implemented for any stochastic project scheduling problem. This phase of the proposed model is comprised of two steps: (i) *%deviation* *analysis of projects* and (ii) *%deviation analysis of activities*. Framework of the uncertainty assessment phase is given in Figure 2.

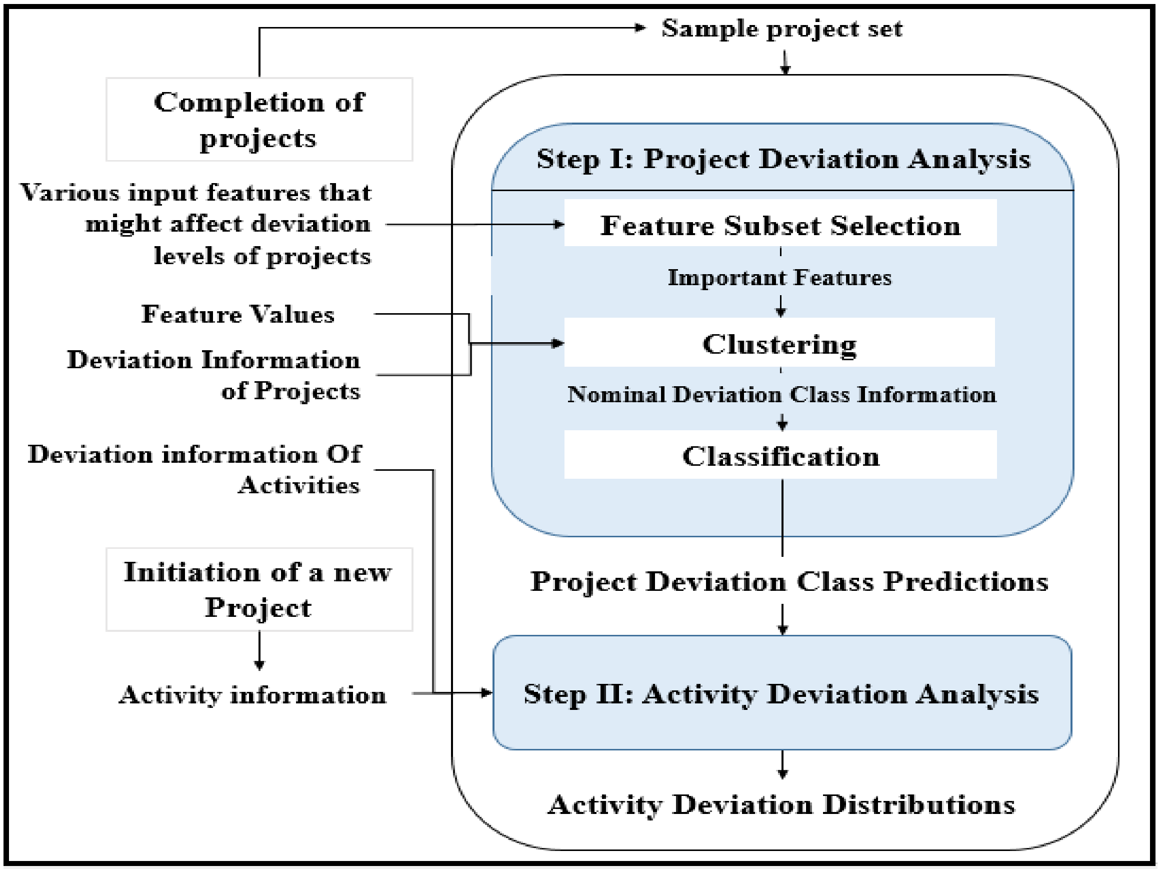

Figure 2. Framework of the uncertainty assessment phase

The details of the proposed uncertainty assessment approach are presented in the following subsections. We kindly suggest the interested readers to refer to Tan et al. (2006), and Du (2010) for detailed information on the data mining tools we use throughout the paper.

4.1. Step I: %Deviation Analysis of Projects

The objective of the first step is to establisha classification model based on completed projects (projects in the sample project set) in the Department in order to classify newly initiated projects with respect to their %deviation levels. The input of this step consists of various features that are thought to be relevant for determining the %deviation levels of projects and the values that these features take for each project. First, the most important features are determined with the help of feature subset selection algorithms, and then clustering is applied to the numerical values of %deviations in order to generate nominal

%deviationclass labels for each project. Afterwards, these nominal and numerical%deviationlevels (outputs) are used in the learning stage of the classification model construction. For each feature subset and output combination, a classification model is constructed and %deviation classes of projects are predicted. All these prediction results are then used to give a probabilistic membership to the projects in the sample project set. Note that when a project is completed in the system, it should be added to the sample data set and Step I should be repeated for better accuracy. The output of this step is various classification models that give probabilistic membership to newly initiated projects. Thus, by using these classification models, in the planning phase, i.e., before the projects actually start, predicting their %deviation levels will be possible and needed precautions can be taken accordingly. Moreover, this stepgives the relations between important features that determine the %deviation level of projects, which enables the project managers to have a better understanding of the system and make fine-tuning on these feature values in order to bring the projects' %deviation to a desired level.

4.1.1. Feature Subset Selection

Construction of the classification model starts with determining the features that can have a positive or negative effect on the %deviation level of projects and finding the best subset of these features in terms of prediction. In our approach, for the feature subset selection, we suggest utilizing an open source data mining software, namely WEKA, developed by Hall et al. (2009), comparing the performances of different feature subset selection algorithms that the software supports and select the best ones in terms of prediction accuracy.

4.1.2. Clustering

The next step after feature subset selection is to cluster the projects in the sample project set to generate nominal %deviation levels that will be used along with numerical %deviation levels. The main reason of this nominal %deviation level determination is that most of the classification algorithms work with nominal output values.

The aim of clustering in general terms isto divide the data set into mutually exclusive groups such that the members of each group are as close as possible to one another, and different groups are as far as possible from one another, where distance is measured with respect to all available features. In this paper, we employ for clustering the *K-means algorithm* developed by MacQueen (1967) to obtain the nominal output values for each project from the numeric output values. The basic idea of the *K*-means algorithm is to divide the data into *K* partitions with an iterative algorithm that starts with randomly selected instances as centers and then assigns each instance to its closest center. Closeness is often measured by the Euclidean distance but can also be measured by some other closeness metric. Given those assignments, the cluster centers are re-determined and each instance is again assigned to its closest center. This procedure is repeated until no instance changes its cluster. The *sum of squared errors* metric is the most preferred evaluation measure of clustering algorithms. For a detailed discussion on clustering methods, we refer to Berkhin (2006) and Jain (2010).

4.1.3. Classification

After obtaining nominal output values, next step is to develop classification models. In that stage, we propose the use of both numeric and nominal output that both represent the %deviations of projects from their mean. In doing so, we will have more than one classification model, one model for each output type-feature subset combination, each having a different performance on the data.

In this step, instead of selecting the classification model that performs best on the given data, we propose using prediction results of several classification algorithms that give reasonable accuracy and produce probabilistic predictions for the %deviation levels of the projects. This is made possible by the various classification algorithms currently available in WEKA (Hall et al., 2009). By doing so, we will be providing probabilistic memberships to the projects in the sample set that represent %deviation level classes. This approach is considered to be more robust than selecting a single classification model and making deterministic predictions accordingly, since providing a probabilistic prediction precludes missing the actual %deviation class of a project and tolerates the error caused by model selection. In fact, instead of making a class prediction, giving a closeness value to each %deviation class is more understandable by the project managers. Thus,this approach makes sense both in terms of convenience of perception and correctness.

4.2. Step II: %Deviation Analysis of Activities

In Step I we develop a model to predict the %deviation level of a newly arrived project based on its various input features. Using this information, in Step II, we develop a model to predict the %deviationlevel of the "activities" of this new project. The aim of Step II is to obtain %deviation distributions for each *project %deviation class* - *activity class* combination to be used in Phase II of the proposed solution approach for robust project scheduling. Therefore, Step II of the uncertainty assessment phase starts with the classification of all project activities, thus forming a number of activity subsets. Forming a distribution requires sufficient number of replications. Since we are dealing with R&D projects and the activities of R&D projects are usually unique with characteristic work contents, such an aggregation and classification is considered to be compulsory.

For each activity class of a newly arrived project, using the %deviation information of already completed activities in the corresponding activity class and the %deviation class prediction of this newly arrived project, we form a %deviation distribution. Note that, the %deviation classes of already completed projects are determined in previous steps, thus we already know the frequency and %deviation level information for the activities in each *project %deviation class* - *activity class* combination. To form the %deviation distribution for an activity class, we set a minimum and maximum value on the %deviationlevel that an activity can take and then this relatively large range is divided into smaller intervals. After that, for each activity class, frequency information for each project class and %deviationinterval is obtained. In this case, since the project's %deviation prediction is probabilistic, we cannot directly use either the frequency distribution for the activity class or the frequency distribution for the *activity class-project %deviation class* combination. We need to adjust the frequency distribution regarding activity classes using the%deviation class of the projects. In each interval, we knowthe number of activities (# activities)completed and the allocation of these activities to the project's% deviation classes. Thus, adjusted frequency information for an interval is obtained by summing the multiplications of # activities in each project %deviation class with the probability of the %deviation level membership of the newly arrived project. As an illustration, assume that we have two %deviationclasses for the projects (*Class1* and *Class2*) and the newly arrived project is predicted to be a member of *Class1* and *Class2* with probabilities 50% each. Also, assume that the %deviationrange in each class is divided into four intervals and there are 40 completed activities in *Class1* and 80 completed activities in *Class2* that lie within the range of the first %deviation interval. Then the adjusted frequency of an activity having a %deviation level in the first interval equals 60(=40 x 0.50 + 80 x 0.5) for this case.

After obtaining these adjusted frequency distributions, the probabilities for an activity having a %deviation level in each interval is calculated and the piecewise linear %deviation distributions are formed for each activity class in the newly arrived project. This distribution is then used to assign %deviation level to the to-be-scheduled activities in Phase II of the proposed two-phase methodology.

Step I of the uncertainty assessment phase is called whenever a project is completed from Phase II and whenever a new project enters the project management system. Step II of the uncertainty assessment phase isused whenever robust project schedules need to be obtained.

5. SAMPLE APPLICATION IN A CASE STUDY

In the implementation of *uncertainty assessment phase* of the proposed two-phase approach for robust R&D project scheduling, R&D project data of a home appliances company is used.The problem environment under consideration contains multiple projects, which are managed with a *stage-gate* approach and most of them are research-based projects that are subject to considerable amount of uncertainty. Human resources in the department are multi-skilled, i.e., each human resource has its own specialty and the degree of that specialty differs from one human resource to another. This makes human resources very critical for the R&D Department since the human resources are not necessarily substitutable. In the remainder of the text, the only resource type considered is human resource. This is due to the relatively high importance of human resource as well as the relatively unrestricted availability of other resources such as laboratory facilities and equipment. This section first introduces the data used in the implementation and its analysis, and then gives the implementation steps of *uncertainty assessment phase* on real data with the findings and results.

5.1. Data Analysis

In the implementation, first a set of completed projects are analyzed and sample project set to be usedin both phases of the proposed two phase approach is determined. Then, relevant input features that might have a positive or negative effect on the %deviation levels of projects are determined and the values all these features take for each project are obtained. After that, the most important features are determined through feature subset selection. Then, the activities of the projects in the project set are classified into six categories to develop a better activity %deviation level prediction procedure for the activities of a newly arrived project. Data section ends with the presentation of the activity data analysis results. Note that in the feature subset selection WEKA (Hall et al, 2009) is utilized.

5.1.1. Determining the Project Set

To determine the sample project set, first a sample project (project *p*) pointed by the R&D department manager of the firm is considered. The reason why project*p* is pointed is that all the *six* resources that are considered to be the most critical ones in terms of the total workloads of them in the R&D Department are the resources that also worked in the activities of p. To obtain the sample project set, all the other projects in which these resources work during the execution of project$p$(during time range tr_1) are filtered from the project database of the firm. Since total number of projects obtained after this filtering process was 117, a reduction process is applied. In this reduction process, the first and last three months of tr_1 are excluded from consideration yielding a new time range tr_2, then a total of 33 projects whose execution time does not lie in tr_2 are removed from the sample project set resulting in a total of 84 remaining projects. From these 84 projects, all the projects starting before 2007 are also removed since the project plans was not detailed enough. Consequently, a project set comprised of 43 interrelated projects in terms of the resources used is obtained.

5.1.2. Determining the Projects' %deviation Class Labels

In order to consider the %deviations of the projects as a risk measure, in the proposed *uncertainty assessment approach*, we suggest utilizing both numeric output (actual %deviation), and nominal output (class labels representing actual %deviation). In the determination of the nominal output labels, the aim is to classify the projects into four %deviation levels (*Negative High %deviation-NHD, Negative Low %deviation-NLD, Positive Low %deviation-PLD and Positive High %deviation-PHD)*. In this approach, first four clusters are obtained through the implementation of the *simple K-Means algorithm* and then labeling is performed based on the resulting clusters. Resulting threshold values in this labeling approach are -%20, %0, and +%25),which are similar to those provided by the R&D Department (-%20, %0, and +%20). Therefore, the results indicate that the projects with %deviation level less than or equal to -%20 are in the class of *NHD*; those with %deviation between -%20 and %0 are in the class of *NLD*; those with %deviation between %0 and +%25are in the class of *PLD*; and those with %deviation more than +%25 are in the class of *PHD*.

5.1.3. Determining the Relevant Features that Affect %deviation Level of Projects

After several interviews with the project managers of the firm, the factors that might affect %deviation level of projects through time and resource overruns and underruns are determined and the values that these features take for each project is obtained. Determined input features and the minimum and maximum values that these features take for the projects in the sample project set are listed in Table 1.

Table 1. Determined Input Features

Feature ID	Feature Name	Type	Min.	Max.
F1	Existence of the technology family *Fluid Dynamics*	binary	0	1
F2	Existence of the technology family *Material Science*	binary	0	1
F3	Existence of the technology family *Thermodynamics*	binary	0	1
F4	Existence of the technology family *Cleaning*	binary	0	1
F5	Existence of the technology family *Vibration and Acoustics*	binary	0	1
F6	Existence of the technology family *Structural Design*	binary	0	1
F7	Existence of the technology family *Power Electronics*	binary	0	1
F8	Existence of the technology family *Electronic Assessment*	binary	0	1
F9	Number of collaborative internal plants	integer	0	5
F10	Number of technology families involved in the project	integer	2	9
F11	Required size of project team in numbers	integer	5	27
F12	Number of required equipment and machine type	integer	0	5
F13	Number of collaborations	integer	0	3
F14	First usage of infrastructure	binary	0	1
F15	Existence of similar projects worked on before	binary	0	1
F16	Planned man-months needed	double	6.1	88.69
F17	Planned equipment-months needed	double	0	119.97
F18	Expected cost of the project	integer	32064	506825
F19	Technology maturity of the project	integer	1	25
F20	Position of the project in the *r&D-R&d* spectrum	integer	1	3

Note that F20 specifies the position of the project in the *r&D - R&d* spectrum where *r&D*represents the solely development-based projects and *R&d* represents solely research-based projects.

5.1.4. Activity Classification

As stated previously, the aim of Step II of the uncertainty assessment phase is to obtain %deviation distributions to be used in Phase II of the robust project scheduling model. Since we are dealing with R&D projects and the activities of R&D projects are unique and the work content is characteristic among all the activities, in order to obtain sufficiently large amount of data for valid activity %deviation distributions, we have categorized the activities of projects in the project set in six activity classes. The list of activity classes determined is as follows:

1. Meeting and Reporting Activity Class

2. Designing, Modeling and Visualizing Activity Class
3. Test, Measurement and Analysis Activity Class
4. Prototyping/Production Activity Class
5. Literature and Patent Search Activity Class
6. Others Activity Class

The classification of the activities is not only based on the work contents but also the density of required resource types (human resource or equipment) of the activities.

5.1.5. Analysis of Data

Before starting the implementation, activity data is analyzed in order to be more familiar with the data In this part, we give some statistics concerning the activities in our project set. While Figure 3 shows the number of projects of each project'%deviation class, Figure 4 shows the number of activities in each project %deviation class.

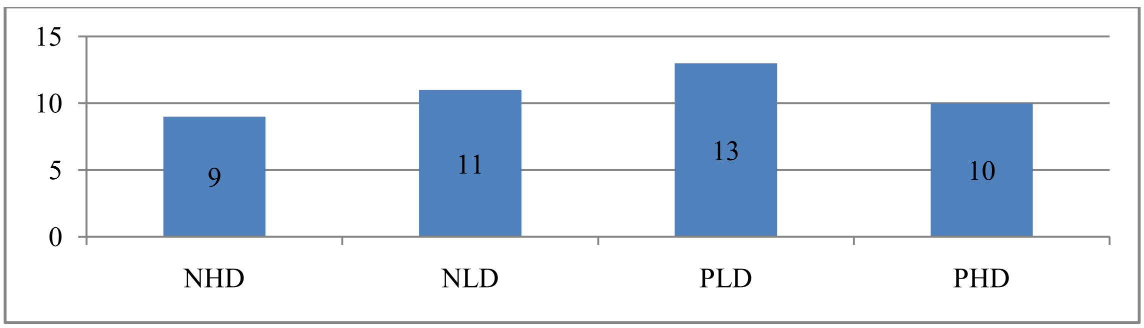

Figure 3. Number of Projects in Each Project %deviation Class

It is seen from Figure 3 that the numbers of projects in each type of project %deviation class are very similar to each other indicating that we almost have a homogeneous project set in terms of project %deviation classes.

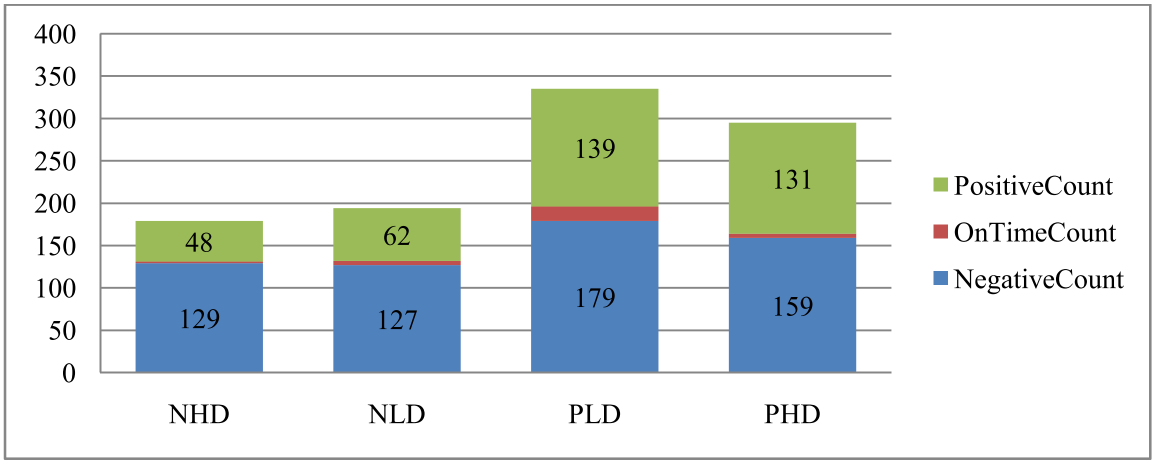

Figure 4. Distribution of Activities in Each Project %deviation Class

Figure 4 indicates while number of activities having negative %deviations is noticeably higher thanthose with positive %deviations in the project classes of *NHD* and *NLD*, this difference is not that much for the project %deviation classes*PLD* and *PHD*.Table 2 shows the number of activities (*#activity*), number of activities with positive and negative deviations (*#positive* and *#negative*) in each activity class along with probabilities of them (*%negative* and *%positive*).

Table 2. Activity Statistics

Activity Class	#activity	#negative	#positive	%negative	%positive
1. Meeting and Reporting	237	134	103	57	43
2. Test Measurement and Analysis	609	364	245	60	40
3. Literature and Patent Search	27	17	10	63	37
4. Design Modeling and	57	33	24	58	42
5. Prototyping/ Production	56	34	22	60	40
6. Others	19	9	10	47	53

Table 3 reveals that the activities in all classes except in “others” have a tendency of having negative %deviation. This shows that the project managers are generally overestimating the resource usages of the activities and adopt a risk-averse strategy not to be blamed in case unforeseen events. Similar analysis is done for the activities in each project %deviation class. Since all the projects dealt with are completed projects, we know the actual project %deviation class of them. Table 3 demonstrates the activity statistics for each %deviation class.

Table 3.Project %deviation Class Based Activity Statistics

Project %deviation	# projec	# activity	# negative	# on-time	# positive	% negative	% on-time	% positive
NHD	9	179	129	2	48	72	1	27
NLD	11	195	127	5	62	65	3	32
PLD	13	338	179	17	139	53	5	41
PHD	10	296	159	5	131	54	0	44

Table 4 indicates that the activities in the projects' %deviation class*NHD* and *NLD* have the tendency of having negative %deviation as expected. On the contrary, the activities in the projects' %deviation class *PHD* and *PLD* have also the tendency of having negative %deviation with a lower probability. This can be explained by the dominance of the activities with negative %deviation in the activity set.

Table 4 shows the number of the activities completed on time, with negative %deviation and with positive %deviation together with their project %deviation classes and the probability of being in these project % deviation classes.

Table 4.Activities' Project %deviation Class Statistics Based on Their %deviation Type

		%activities					
	#projects	NHD	NLD	Total %negative	PLD	PHD	Total %positive
%negative	**594**	22	21	43	30	27	57
%on-time	**29**	7	17	24	59	17	76
%positive	**380**	13	16	29	37	34	71

Table 4 reveals that among the activities having negative %deviation, about half of them are in the class of *NHD* and *NLD*, among the activities on time, 75% of them belongs to the class of *NLD* and *PLD*, and among the activities having positive %deviation 71% of them belongs to the class of *PLD* and *PHD*.

5.2. Step I: Projects'%Deviation Analysis

5.2.1. Feature Subset Selection

Not only missing some of the significant input features but also the existence of ample number of irrelevant features makes it difficult to establish the relation between the inputs and the output. Therefore, *feature subset selection* is an essential step in data mining process and directly influences the classification performance. In the analysis reported here, determined 20 input features (FS_1) and 11 input features (FS_2) resulting in exclusion of the features regarding to existence of various technology families in the projects are utilized with two types of numeric output: *percentage human resource %deviations of the projects* (*%deviation*) and *absolute percentage human resource %deviations of the projects* (|*%deviation*|). Various different feature subset selection algorithms supported in WEKA are utilized. In these algorithms, a subset of the data instances are used as a training set, while the remaining instances are used as test instances to evaluate the performance of the algorithm on these test instances. Note that different training and test combinations yield different subsets of significant inputs hence a threshold value of 70% is set in order to make a final decision for inclusion of a feature for the further analysis. The results of the feature subset selection analysis are given in Table 5.

Table 5. Results of Feature Subset Selection Analysis

Feature Subset	**Results for FS_1**		**Results for FS_2**	
Output Type	*%deviation*	|*%deviation*|	*%deviation*	|*%deviation*|
Selected Feature	F1, F6, **F13**, **F14, F15**	F1, F4, F8, **F14, F15**	F10, **F13**, **F14, F15**, F19	F10, F11, **F13, F14, F15**

As a result of the analysis, four different feature subsets are determined as significant: {F1, F6, F13, F14, F15} ($\boldsymbol{FS_3}$), {F1, F4, F8, F14, F15} ($\boldsymbol{FS_4}$), {F10, F13, F14, F15} ($\boldsymbol{FS_5}$), and {F10, F11, F13, F14, F15} ($\boldsymbol{FS_6}$). Notice that F14 and F15 (*First usage of infrastructure* and *Existence of similar projects worked on before*) is included in all determined feature subsets which indicates the high importance of the experience level of the resources on the subject of the projects. We also see that F13 (*Number of collaborations*) is included in three of the determined feature subsets and when FS_1 is (*Effect of various technology families' existence*) considered, F1 replaces F10 indicating the relative importance of existence of the technology family *Fluid Dynamics* over the number of total technology families involved. All these results enable the project managers to see the relations of these various features between each other and on the expected performance of the project. In order to evaluate the influence of the feature subset selection stage to the classification performance, two additional feature sets are also included in further analysis: FS_1 and FS_2.

5.2.2.Classification

In this step, for each feature subsetconsidered in the previous subsection, classification analysis is performed on both numeric and nominal output values.

Table 6. Classification Results for Numeric Output

Method	Perf. Metric	Input Feature Subset						
		FS_1	FS_2	FS_3	FS_4	FS_5	FS_6	AVG.
LR	*%Correct*	21%	49%	40%	37%	40%	28%	36%
	MSE	94	54	57	63	64	66	66
	Used Features	F1,F4,F6,F11, F13,F16,F18,F9	F11,F13, F14,F19	F10,F14, F19	F11,F13	F1,F6,F13	F1,F4	
LMSLR	*%Correct*	23%	40%	40%	37%	47%	56%	40%
	MSE	93	34	37	47	42	34	48
	Used Features	ALL	ALL	ALL	ALL	ALL	ALL	
PR	*%Correct*	49%	40%	50%	37%	42%	51%	45%
	MSE	37	44	36	45	42	39	41
	Used Features	F1,F4,F6,F11, F13,F9,F19	F10,F11,F13, F14,F19,F20	F10,F13, F14,F19	ALL	ALL	ALL	
M5P	*%Correct*	56%	47%	44%	37%	40%	28%	42%
	MSE	49	35	36	45	64	55	47
	Used Features	F6,F11,F13,F19	F11,F13,F16, F19,F20	F10,F13, F14,F19	F11,F133	F1,F6,F13	F1,F4	AVG.

Classification with Numeric Output

In classification with numeric output, only regression based classification algorithms supported by WEKA are used: *Linear Regression (LR)*, *Least Median Squared Linear Regression (LMSLR)*, *Pace Regression (PR)* and *M5P Algorithm*. Table 6 shows the predictive performance of these algorithms based on two metrics (*Accuracy Rate (%Correct)* and the *Mean Squared Error (MSE)*) for each of the six input feature subsets determined previously. Note that, for the numerical output case *#Correct* are based on the previously determined nominal output labels (*NHD, NLD, PLD*, and *PHD)*. In order to calculate the *MSE* of classification methods number-based labels as 1, 2, 3, and 4 corresponding *NHD, NLD, PLD*, and *PHD,* respectively are used. Thus, the error is simply calculated as the difference between the corresponding number-based prediction label and number-based label. Table 6 also presents the features the performance results for each feature subset used in the analysis.

Notice that, since the classification algorithms have embedded feature selection mechanisms, Table 6 also includes a row that shows the selected features in the implementation of the mentioned classification algorithms. The results show that the best average *%Correct* and MSE values are obtained with *PR*.

Classification Analysis with Nominal Output

The classification algorithms applied to the data set with nominal output were *J48 Decision Tree (J48) classification method and Naive Bayes (NB*) classification method. Again the same predictive performance metrics are used. The results for the data set with nominal output are presented in Table 7.

Table 7. Classification Results for Nominal Output

Method	**Perf. Metric**	**Input Feature Subset**						
		FS_1	FS_2	FS_3	FS_4	FS_5	FS_6	**AVG.**
J48	***%Correct***	84%	84%	65%	67%	63%	50%	69%
	MSE	20	12	30	20	41	45	28
	Used Features	F1,F3,F4,F5, F6,F12,F13,F14, F15,F17,F20	F11,F12,F13, F14,F18,F19	F10,F11, F13,F14	F1,F1, F14	F1,F13, F14	F1,F4, F8,F14	
NB	***%Correct***	67%	60%	54%	54%	56%	50%	57%
	MSE	22	29	44	38	33	45	35

Table 7 shows that on average J48 and NB, which work on nominal output values, generate better accurate results than the regression based methods. Moreover, best average *%Correct* and MSE values are obtained with *J48 Decision Tree Method*. Notice that when all the features determined are used, accuracy rate rises up to 84%.

Further Results

In this part, we suggest two further ways of producing classification results. In addition tooption of selecting a feature subset and a classification method, which gives the best accuracy (*J48* with FS_1), another option is to use all the analysis done so far and producing probabilistic results. Using the prediction results obtained with each feature subset and classification model combination, we can provide probabilistic %deviation estimations for each project by simply counting each label assigned to projects and dividing this number to the number of prediction methods. For the numeric and nominal %deviation, we have 36 predictionsin total for each project. By using probabilistic results the aim is to decrease the prediction error arising from the selected feature subset and classification method. The probabilistic results for a subset of projects are shown in Table 8.

Table 8. Probabilistic Classification Results for Numeric Output

	Prediction Count				Probability			
Project ID	NHD	NLD	PLD	PHD	NHD	NLD	PLD	PHD
10-015	28	8	0	0	78%	22%	0%	0%
09-045	24	12	0	0	67%	33%	0%	0%
08-054	15	15	6	0	42%	42%	17%	0%
09-018	9	8	10	9	25%	22%	28%	25%
09-023	13	18	5	0	36%	50%	14%	0%
09-036	4	20	12	0	11%	56%	33%	0%
11-009	5	12	16	3	14%	33%	44%	8%
10-049	9	15	12	0	25%	42%	33%	0%
08-040	3	21	7	5	8%	58%	19%	14%
08-022	3	18	11	4	8%	50%	31%	11%

Table 9 shows the performances of classification analysis for prediction of the classes using probabilistic results in terms of the number of projects estimated to be in *NLD* and *NHD*, in *PLD* and *PHD* and in terms of the number of exact class prediction.

Table 9. Performances of Probabilistic Results for the %deviation Level Prediction

Output Type	**#negative match**	**#positive match**	**#exact match**	**%negative match**	**%positive match**	**%exact match**
Numerical	15	16	22	35%	37%	51%
Nominal	19	13	29	44%	30%	67%

Table 9 shows that the proposed probabilistic approach performs well for the prediction of %deviation level classes of projects but performs much betterwhen only negative and positive class labels are considered together, i.e., only two classes are considered as negative %deviation class and positive %deviation class. Table 9 also shows that compared classification predictions made with nominal output provide better results.

Another way of prediction of the %deviation levels of projects could be adopting a one-take-out procedure. In this iterative one-take-out procedure, in each iteration one project is disregarded from the analysis and the learning stage of the classification algorithm is performed from the remaining 42 projectsresulting accuracy for each project. Accuracy of the method is then the average accuracy of the results for all projects.

<u>*5.2.3. Comparisons of Classification Approaches*</u>

In the previous sub-sections we have provided the classification accuracy results using each of output and feature subset combinations. To make a better decision on selecting the classification approach, considering solely accuracy results and selecting the method giving the best accuracy might not be reliable enough. In this part, we will suggest further ways of comparing classification methods used in the previous sections.

One way of comparing classification approaches other than comparing accuracy performance is the useof average variability of each classification approach. This variability attribute is specific to each *feature subset - classification method* combination and can be calculated using the number-based labelsassociated with the projects used in calculationof *MSE*. The variability of a project for *feature subset - classification method* combination is simply the sum of the squared differences between the predicted number-based label for the combination in question and predicted number-based labels obtained for the remaining *(feature subset - classification method)* combinations. The average variability of a combination is then obtained summing these variability values for all projects and simply taking the average. Since the number of combinations for each output type is different (due to number of algorithms used in the analysis for the corresponding output type) in order to make the comparisons consistent we have also divided the average variability values to the number of combinations. In doing so, we were able to compare the (*feature subset - classification method*) combinations between each other. Results showing the average variability of the combinations are demonstrated in Table 10.

Table 10. Average Variability Results of the Classification Approaches

Feature Subset	Classification Method	Average Variability	Feature Subset	Classification Method	Average Variability
FS_1	*LR*	1.76	FS_4	*LR*	1.01
	LMSLR	1.89		*LMSLR*	0.78
	PR	0.73		*PR*	0.90
	MP5	1.00		*MP5*	1.01
	J48	0.71		*J48*	0.66
	NB	0.86		*NB*	0.59
FS_2	*LR*	0.94	FS_5	*LR*	1.80
	LMSLR	0.94		*LMSLR*	0.99
	PR	0.80		*PR*	0.97
	MP5	0.73		*MP5*	1.80
	J48	0.64		*J48*	1.02
	NB	0.86		*NB*	0.52
FS_3	*LR*	1.14	FS_6	*LR*	1.26
	LMSLR	0.78		*LMSLR*	0.76
	PR	1.02		*PR*	0.88
	MP5	0.98		*MP5*	1.48
	J48	0.53		*J48*	0.69
	NB	0.65		*NB*	0.70

Table 10 indicates that among the classification methods PR, J48 and NB give the lowest average variability results for FS_1, FS_2 andFS_3, and FS_4, FS_5 and FS_6; respectively. In parallel with the accuracy results, the use of classification methods that usenominal labels obtained by applying the *simple K-Means clustering algorithm*, obtains better variability values for all feature subsets.

Another consideration we need to take into account while comparing classification approaches is the interpretability of the results. Since the Naive Bayes classification method is a black box only giving the classes of the given projects, it might be hard to convince the decision-maker about the reliability of the method. Decision tree based algorithms are better for interpretability, since they also provide the tree as a rule of classification to the decision maker for the newly added data point (i.e., a new project in our case). When selecting a classification approach, another consideration is the number of features used in the classification and the ease of obtaining them.

5.3. Step II: Activities'%deviationAnalysis

In Step I, we develop a model to predict %deviation level of a newly arrived project based on its various input features. Using this information along with activity class information in Step II,we develop a model to predict %deviation of the activities of this newly arrived project. In this section, first, we present activities' %deviation analysis results for a project whose %deviation class is deterministically predicted, and then for a project whose project %deviation class is probabilistically predicted.

5.3.1. Activity %deviationPrediction with Deterministic Project %deviation Class Prediction

Using the classification model that gives the best accuracy rate, for a newly arrived project we predict its %deviation class and for each activity class in the corresponding project using the %deviations of already completed activities in the associated activity class we form a %deviation distribution. To illustrate this distribution forming process, Table 11 shows the frequency information for *NHD Project Class - Test, Measurement and Analysis Activity Class* combination and Figure 5 shows the chart of the corresponding %deviation distribution.

Table 11. Frequency and Probability Information for the NHD-Test, Measurement and Analysis Class Combination

Activity Class	#activities in *NHD* Project %deviation Class	%deviation Range	#activities	Probability of Being in the Range
2. Test Measurement and Analysis	102	(-1)-(-0.67)	23	22.55%
		(-0.67)-(-0.33)	30	29.41%
		(-0.33)-(0)	23	22.55%
		0-0.33	14	13.73%
		0.33-0.67	6	5.88%
		0.67-1	1	0.98%
		1-1.33	3	2.94%
		1.33-1.66	0	0.00%
		1.66-2	2	1.96%

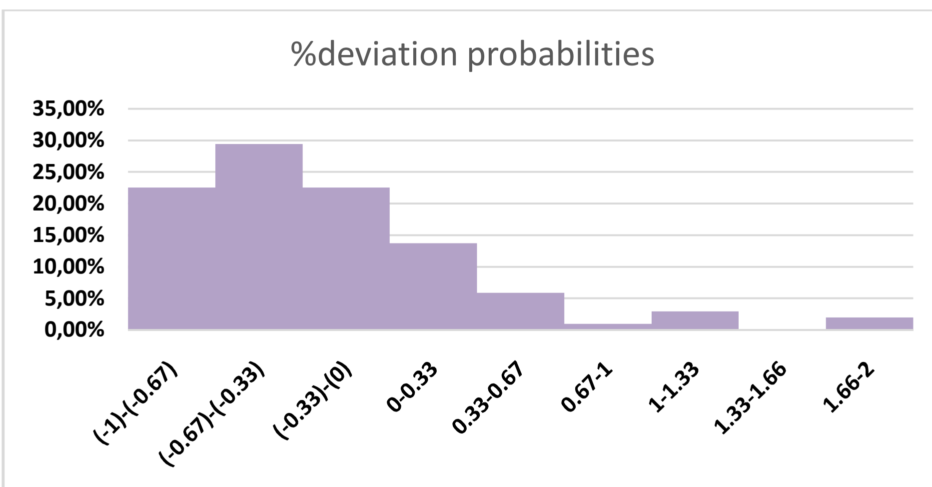

Figure 5. Distribution the *NHD - Test, Measurement and Analysis* Combination

%deviation distributions of the activities belonging to each *activity class - project %deviation class* combinations are obtained following the same procedure and predictions are made for all the activities belonging to the projects in the sample project set in order to make comparisonswith the actual %deviation levels.

5.3.2. Activity %deviation Prediction with Probabilistic Project %deviation Class Prediction

In this section, the procedure is modified using the probabilistic results obtained in Step I. In doing so, we do not ignore the possibility of missing the exact %deviation level of projects. To do so, for each activity class, adjusted frequency information is used. Table 12 tabulates the frequency information for the activity class "Meeting and Reporting" and Figure 6 illustrates the corresponding frequency chart.

Table 12. Frequency Information for the Activity Class "Meeting and Reporting"

Activity Class	#activity	%deviation range	*NHD*	*NLD*	*PLD*	*PHD*
1. Meeting and Reporting	187	(-1)-(-0.67)	13	16	7	18
		(-0.67)-(-0.33)	5	5	6	4
		(-0.33)-(0)	10	9	8	6
		0-0.33	6	11	6	7
		0.33-0.67	1	4	4	2
		0.67-1	0	2	6	5
		1-1.33	1	0	2	1
		1.33-1.66	0	1	2	1
		1.66-2	2	2	6	5

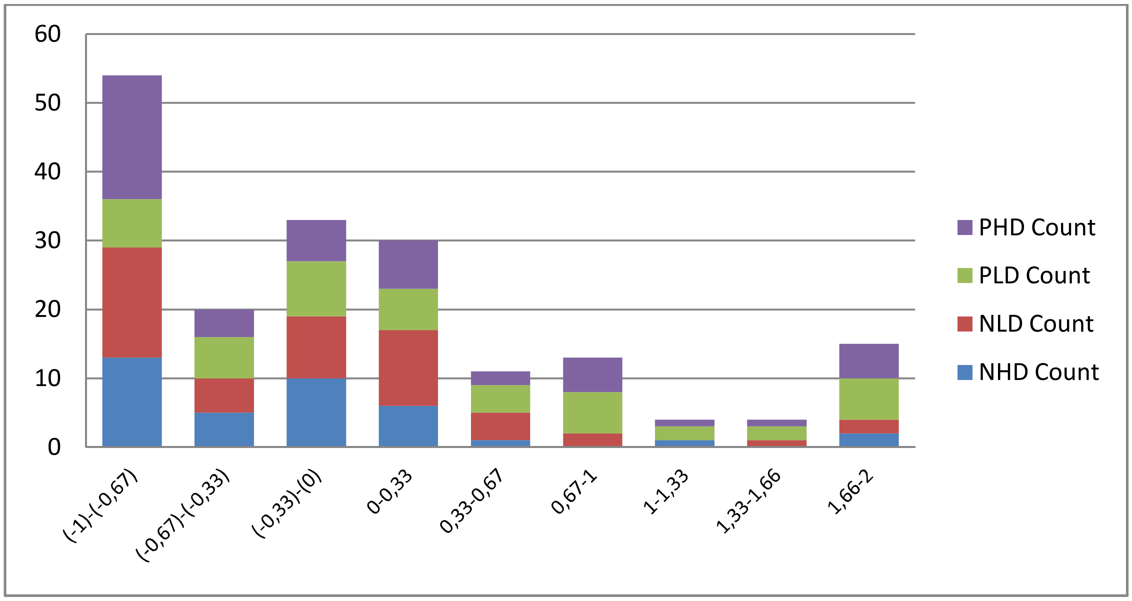

Figure 6. Frequency Distribution for Activity Class "Meeting and Reporting"

In this case, since the project's %deviation prediction is probabilistic, we cannot directly use frequency information, we need to adjust it by multiplying the frequency probabilities in each interval with the membership probabilities. Figure 7 shows the %deviation distribution for the activities belonging to *Test and Measurement* activity class for a project with a membership 50% each for PHD and PLD project %deviation classes.

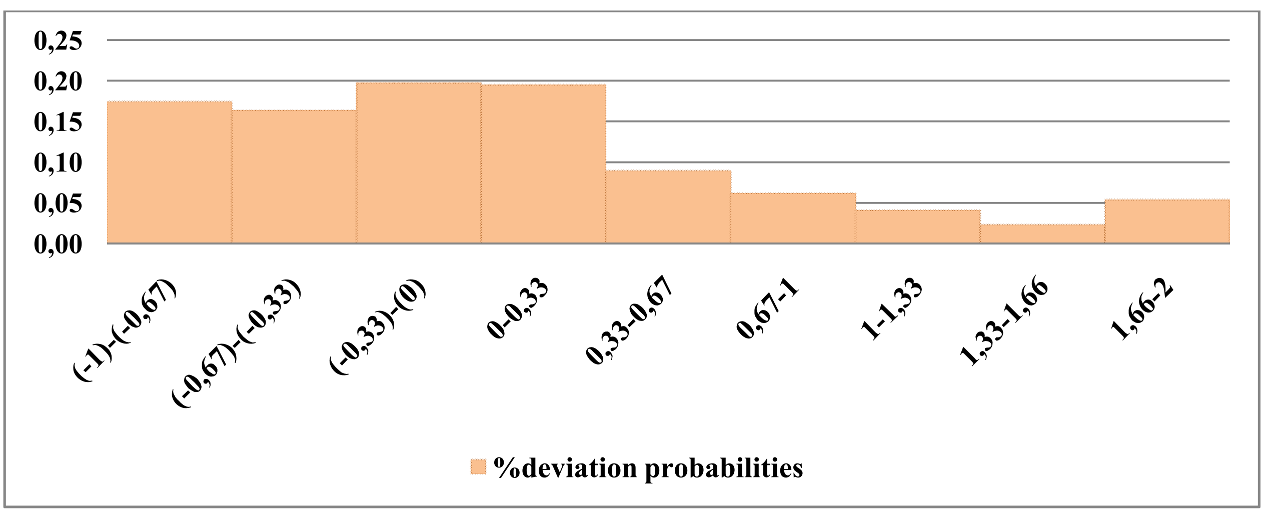

Figure 7. Probability Distribution for *Test and Measurement Activity Class- 50% PHD, 50% PLD Project %deviation Class* Combination

5.3.3. Activities'%deviationPrediction Performance Analysis

To test our activity %deviation prediction procedure we have used the actual project %deviation class information and actual activity %deviations. The performance is measured in terms number of matches for activities having negative %deviation, and number of matches for activities having positive %deviation. To reduce the effect of randomness, we performed five replications. While Table 13 shows the results for the actual %deviation labels are used for projects, Table 14 and 15 show the results when projects' %deviation levels are predicted deterministically and probabilistically.

Table 13. %deviationPrediction Results for the Actual Project %deviation Classes

	Prediction Replication					**AVG.**
	1	2	3	4	5	
Total #negative Match	377	366	373	365	353	366.8
Total #positive Match	168	161	183	153	146	162.2
Total #activity Match	545	516	556	518	498	526.6
% Negative Match	60.03%	58.28%	59.39%	58.12%	56.21%	58.41%
%Positive Match	44.21%	42.37%	48.16%	40.26%	38.42%	42.68%
%Match	54.07%	51.19%	55.16%	51.39%	49.40%	**52.24%**
Total #activity						1008
Total #activity Having Negative %deviation						628
Total #activity Having Positive %deviation						380

Table 14 indicates that using the procedure that we suggested with deterministic project %deviation level prediction, we are able to make correct predictions on the %deviations of activities on the average with a probability of 51%. Our predictions are much better to predict the negative %deviations of activities than the positive %deviations of activities. Table 15 shows that using the procedure that we suggested with probabilistic project %deviation level predictions; we are able to make correct predictions on the %deviations of activities on the average with a probability of 52%. Similarly, our predictions are much better to predict the negative %deviations than the positive %deviations for activities. Notice that the performance of the proposed prediction procedures are almost the same when the deviation level of projects are exactly known in advance.

Table 14. %deviationPrediction Results with Deterministic Project %deviation Class Prediction

	Prediction Replication					**AVG.**
	1	**2**	**3**	**4**	**5**	
Total #negative Match	369	371	375	360	369	368.8
Total #positive Match	147	143	160	160	140	150
Total #activity Match	516	514	535	520	509	518.8
% Negative Match	58.76%	59.08%	59.71%	57.32%	58.76%	58.73%
%Positive Match	38.68%	37.63%	42.11%	42.11%	36.84%	39.47%
%Match	51.19%	50.99%	53.08%	51.59%	50.50%	**51.47%**
Total #activity					1008	
Total #activity Having Negative %deviation					628	
Total #activity Having Positive %deviation					380	

Table 15. %deviationPrediction Results with Probabilistic Project %deviation Class Prediction

	Prediction Replication					**AVG.**
	1	2	3	4	5	
Total #negative Match	375	362	360	353	379	365.8
Total #positive Match	160	155	156	159	151	156.2
Total #activity Match	535	517	516	512	530	522
% Negative Match	59.71%	57.64%	57.32%	56.21%	60.35%	58.25%
%Positive Match	42.11%	40.79%	41.05%	41.84%	39.74%	41.11%
%Match	53.08%	51.29%	51.19%	50.79%	52.58%	**51.79%**
Total #activity					1008	
Total #activity Having Negative %deviation					628	
Total #activity Having Positive %deviation					380	

To illustrate how the activity %deviation prediction procedure will work for two different projects having different %deviation prediction profiles, the probabilities of activities' tendency of having negative and positive %deviations are illustrated in Table 16.

Table 16. The Probabilities of Activities' Having Negative and Positive %deviations for Two Different Projects

	Project Type (0.00, 0.10, 0.10, 0.80)*		Project Type (0.80, 0.10, 0.10, 0.00)*	
Activity Class	**Negative Probability**	**Positive Probability**	**Negative Probability**	**Positive Probability**
1. Meeting and Reporting	56.24%	43.76%	68.58%	31.42%
2. Test Measurement and Analysis	54.35%	45.65%	71.14%	28.86%
3. Literature and Patent Search	68.57%	31.43%	53.33%	46.67%
4. Design Modeling and	40.23%	59.77%	54.97%	45.03%
5. Prototyping/Production	56.25%	43.75%	75.00%	25.00%
6. Others	56.00%	44.00%	40.00%	60.00%
**(%NHD, %NLD, %PLD, %PHD)*				

Table 16 shows that the activities in the *Meeting and Reporting*, *Test Measurement and Analysis*, *Literature and Patent Search*, *Prototyping and Production,* and *Others* activity classes belonging to mostly to *PHD* project class and mostly to *NHD* project class tend to have negative %deviation. On the other hand, while the activities in the *Design Modeling and Visualizing* activity class belonging mostlyto *PHD* project class tend to have positive %deviation, the ones belonging mostly to *NHD* class have the tendency of having negative %deviation.

6. SUMMARY AND CONCLUSIONS

In this study, we presentthe *uncertainty assessment phase* of the proposed two-phase approach for robust project scheduling. Using the *feature subset selection*, *clustering* and *classification* tools, we develop a two-step uncertainty assessment model to be used to predict the deviation levels of projects and their activities. While in Step I we develop classification models to predict the deviation level of a newly arrived project, in Step II we develop an activity deviation prediction procedure for the activities of a newly arrived project. Step I enables the project managers to predict the deviation level of a project before it actually starts and also enables the project managers to take the needed precautions by detecting the risky projects. Step II not only constitutes the input for proactive project scheduling to obtain robust baseline project schedules by presenting activity deviation distributions but also identifies the risky activities that need close monitoring. In doing so, project managers may focus their attention to the risky activities that are expected to have high uncertainty levels and take the needed precautions to bring their deviation levels to a desired level by adding additional resources and/or by monitoring the progress of these activities more closely.Especially in a multi-project environment, all these supportive features of the proposed uncertainty assessment model helps the project managers to make more analytical and comprehensive decisions when managing projects.

We also present a real case application of the uncertainty assessment model using the R&D project data of a leading home appliances company. We make predictions on the %deviation of activities that we have in the sample project set. Besides this main output, we determine the main features that have an effect on the deviations level of projects, and develop classification models to classify the newly arrived projects with respect to their deviation level. The results show that, our uncertainty assessment model works well on the sample project set.

It should also be noted that, proposed models are appropriate not only for the specific case of the resource constrained project scheduling problem, which was introduced in this paper, but also for all project management environments with considerable uncertainty.

REFERENCES

Berkhin, P. (2006). A survey of clustering data mining techniques. In: *Grouping Multidimensional Data* (pp. 25–71). Springer Verlag, Berlin.

Creemers, S., Demeulemeester, E., Van de Vonder, S. (2011). Project risk management: A new approach. In: *2011 IEEE International Conference on Industrial Engineering and Engineering Management* (IEEM), (pp. 839-843).

Demeulemeester, E., Herroelen, W., 2002. *Project Scheduling – A Research Handbook*, International Series in Operations Research and Management Science, vol. 49. Boston: Kluwer Academic Publishers.

Du, H. (2010). Data Mining Techniques and Applications: An Introduction. United Kingdom, UK: Course Technology Cengage Learning.

Project Management Institute. (2013). *A Guide to the Project Management Body of Knowledge* (PMBOK® Guide).Newtown, PA: Project Management Institute, Incorporated.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.*,11(1), 10–18.

Hartmann, S., & Kolisch, R. (2000). Experimental evaluation of state-of-the-art heuristics for the resource-constrained project scheduling problem. *European Journal of Operational Research*, 127(2), 394–407.

Herroelen, W. (2014). A risk integrated methodology for project planning under uncertainty. In: P.S. Pulat (eds.), *Essays in Production, Project Planning and Scheduling* (pp. 203–217). International Series in Operations Research & Management Science 200, Springer Science+Business Media. New York.

Hulett, D. (2009). *Practical Schedule Risk Analysis*. Gower Publishing, Ltd.

Jaafari, A. (2001). Management of risks, uncertainties and opportunities on projects: time for a fundamental shift. *International Journal of Project Management*, 19(2), 89–101.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651-666.

Kirytopoulos, K., Leopoulos, V., Malandrakis, C., & Tatsiopoulos, I. (2001). Innovative knowledge system for project risk management. *Advances in Systems Science and Applications*, 505–510.

Leus, R., & Herroelen, W. (2004). Stability and resource allocation in project planning. *IIE Transactions*, 36(7), 667–682.

MacQueen, J., & Others. (1967). Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium On Mathematical Statistics and Probability* (Vol. 1, pp. 281–297).

Schatteman, D., & Herroelen, W. (2008). Methodology for integrated risk management and proactive scheduling of construction projects. *Journal of Construction Engineering and Management*, 134(11), 885–893.

Tan, P.-N., Steinbach, M., Kumar, V. (2006). Introduction to Data Mining (Vol. 1). Pearson Addison Wesley, Boston.

Van de Vonder, S., Demeulemeester, E., & Herroelen, W. (2007). A classification of predictive-reactive project scheduling procedures. *Journal of Scheduling*, 10(3), 195–207.