Modern Applied Science; Vol. 8, No. 6; 2014
ISSN 1913-1844 E-ISSN 1913-1852
Published by Canadian Center of Science and Education

Estimation of Project Completion Time-Based on a Mixture of Expert in an Interactive Space

M. T. Hajiali[1], M. R. Mosavi[2] & K. Shahanaghi[3]

[1] Ph.D. Student, Department of Industrial Engineering, Iran University of Science and Technology, Narmak, Tehran, Iran

[2] Professor, Department of Electrical Engineering, Iran University of Science and Technology, Narmak, Tehran, Iran

[3] Assistant Professor, Department of Industrial Engineering, Iran University of Science and Technology, Narmak, Tehran, Iran

Correspondence: M. T. Hajiali, Department of Industrial Engineering, Iran University of Science and Technology, Narmak, Tehran 16846-13114, Iran. E-mail: hajialinajar@iust.ac.ir

Received: September 4, 2014 Accepted: October 8, 2014 Online Published: November 8, 2014

doi:10.5539/mas.v8n6p229 URL: http://dx.doi.org/10.5539/mas.v8n6p229

Abstract

Estimation the time and the cost of completing projects on the basis of decision making to use either of the estimation methods are one of the most important issues in project management. In this paper, a decision making database of learning machines, is proposed, that a set of possible estimator are working together to estimate the project completion time, in it. This cooperation is based on samples neighborhood in the feature space. One of the important issues, that learning machines are facing it, is the complexity in feature space, because of features with high-correlation. In this paper, to avoid this problem, principal component analysis (PCA) method is used to accuracy has increase, addition to, increasing in system speed. Moreover, methods based on the ensemble, have a higher reliability and ability to generalization, compared to single methods. Furthermore, the hybrid method, (PCA and ensemble), have all the above mentioned advantages. Therefore, system reliability control, using more powerful learning machines, in ensemble, and also ability of the proposed model, to manage existing poor estimators, in ensemble, are other important features of this method. In the end, a software code was created, which provides ability to connect to MSP.

Keywords: Principal Component Analysis (PCA), system reliability control, project management, decision making

1. Introduction

One of the most effective methods in predicting the time and the cost of project completion is use of earned value management. Formula-based methods were used frequently in the past, but nowadays, with increasing volume of received information, from projects, an increasing need to more powerful and more reliable methods, is felt. The methods based on data mining, gradually have widely been used as one of the important methods in this field. Among these methods, can refer to the methods based on neural network, which are used in [1]; in this paper, a neural network with 5 inputs and 5 outputs, and also a hidden layer, trained to predict the actual cost. The process for the minimization of the error is repeated, step by step, until a termination criterion for the system is achieved. Methods based on time series, also have been able to provide, acceptable results in this field, especially, high ability of these methods in predicting future events is the most important factor of these methods [2, 3].Methods based on S shape curves are the other methods, which be extensively used in this field [4, 5, 6].In paper [6], using S shape curves, and the relationship between cash flow and project progress; linear and nonlinear models of project progress, taking into consideration uncertainties in cash flow**.** In paper [7], a dynamic model obtained from earn value management (EVM), using ARMA from the project progress, and then, in an uncertainty space, time and cost of completion of the work, estimated using Kalman fuzzy method.

Support Vector Machine (SVM) is type of learning machines, which has a very high ability in modeling datasets, with a highly complex feature space. Hence, several methods based on SVM have been proposed for estimating the time and cost of projects, which proposed in[8, 9, 10]. In paper [8], a combination of genetic algorithm and

SVM has been proposed. Also in [9], a combination of PCA and SVM is used to predict the project time completion. In paper [10], weighted SVM (wSVM), which is an improved model of SVM, is used to reduce the effects of noise and irrelevant data. This model was used with fuzzy logic and genetic algorithms, as fmGA, to calculate Estimate At Completion (EAC).In this article, methods based on fuzzy approach, such as uncertainty in the obtained value calculations, which inputs and outputs of the system, have considered as a fuzzy variable.

In papers [11, 12, 13], methods based on EVM, reviewed, and new indicators are provided for work in this field. In paper [14], these three methods are compared with each other, and also the efficiency of cost classic indicators, such as SV and SPI, are compared with the efficiency of new indicators, which are similar to the $SV\ (t)$ and $SPI\ (t)$, in units of time. In this paper, we have used 14 different features, which were used for EVM, in paper [14]. In Table 1, some features used in these methods are listed.

Table 1. Features used in this paper

AC	Actual cost	SCI	Schedule cost index
EV	Earned Value	AD	Actual duration
PV	Planned value	ES	Earned schedule
SV	Schedule variance	PD	Planned duration
CV	Cost variance	SV(t)	Schedule variance time
SPI	Schedule performance Index	SPI(t)	Schedule performance index time
CPI	Cost performance index	SCI(t)	Critical ratio time

It should be noted that features used in this articlemostly have a high correlation in feature space; therefore in this paper a combination of Principal Component Analysis (PCA) and ensemble methods, is used.Methods based on ensemble, have attracted a lot of attention to themselves in various fields, such as image processing and computer vision, and credit scoring [15-19].Also in mentioned papers, only one estimators used for estimating the project. Using only one estimator, in learning process can be problematic. Due to the dynamic nature of the problem, in project management, tuninga learning machine on a specific Dataset, however, can produce good results, but it is very risky for a practical application and has low reliability. Methods, based on ensemble have higher reliability and generalization, because of increasing diversity in decision making process, and also combination of these methods with PCA algorithm, has been able to produce much better results than individual learning machines in ensemble.

2. Proposed Method

In data mining methods such as classification and regression; first, available data sets should be split into two parts of training set and test set.In the proposed method, how to divide the data set is, in a way that previous carry out projects placed in the training set. It helps to increase knowledge from the problem, and produce a better approximation and also is suitable to the simulation of dynamic project. By developing the current project in each step, produced knowledge is added to the training set, and the rest of current project place in the test set.For the preparation of data collection, first, features are normalized. After normalization of the dataset, one of the crucial problems, that prediction faces it, is the curse of dimensionality phenomena. This problem makes the feature space very complex. Existence characteristics that have a strong correlation together, make the forecast slow, and also can have a negative impact on the accuracy of the proposed method. As it has been mentioned before, EVM features have strong linear relationships with each other, so using all the features can have a negative impact on the forecast. In this paper, we tried to solve the problem using a well-known technique, PCA.

After simplification dataset, and extracting principal components, dataset, is ready to implement the proposed method. In this paper, an ensemble method is used; in which four different estimators interact with each other. This method can be called as a method of ensemble of expert that final result is produced based on a fusion method. In this paper, the similarity between neighboring samples is used in the feature space to different estimators have a collaboration together. As we know, the principle of neighborhood is one of the important issues in the data mining, and in other areas, such as Image processing. Different methods such as K-nearest neighbor (KNN), attracted a lot attention for different applications.

Hence, the process is as follows that initially, all estimators be trained on the training set at every stage of the project. As a result, the estimator learns the learning sets and be prepared to estimate results for data from the test set. This work is performed in this way that, initially, the KNN algorithm run on each data sample in test set. Based on terms of user K-nearest neighbors are found for each sample of the test, of the training set, then, the

neighborhoods ordered as ascending. The nearest neighbors have most closely resembles and farthest neighbors have least similarity to the test sample. Accuracy of learning machines on a single training example, in the previous step is known, according to this, we can identify the best estimator for different K neighbors. Suppose, if among the three existing learning machines, in ensemble, 2 learning machines have the best result on the 5 neighborhoods, so both two learning machines will be used to estimate the result of the test sample. In conclusion, these results should be fused, together. To this work, a weighted function based on distance between samples and test sample in the feature space was proposed. Equation 1 shows how calculation of estimated amount for each sample. Where, $PE\ (t)$ is the predicted amount from proposed ensemble method, on sample t, and, $Pi\ (t)$ is estimator i, from estimators that have best results on the K neighbors, and Wi is the weight.$PE\ (t)$ is defined in equation 1.

$$PE(t) = \sum_{i=1}^{N} w_i P_i(t) \tag{1}$$

It should be noted that weight has an inverse relationship with distance, hence, should be reduced by increasing the sample distance, also, for the weight should we have: $\sum_{i=1}^{K} w_i = 1$. In this paper, first, weights are sorted in descending order, then, the relationship between weights calculated as follow, where $Dist$ is an array of distances. A proposed weight is defined in equation 2.

$$w_i = \frac{Dist(i)}{sum(Dist)} \tag{2}$$

Various stages of the proposed method are shown in Fig 1.

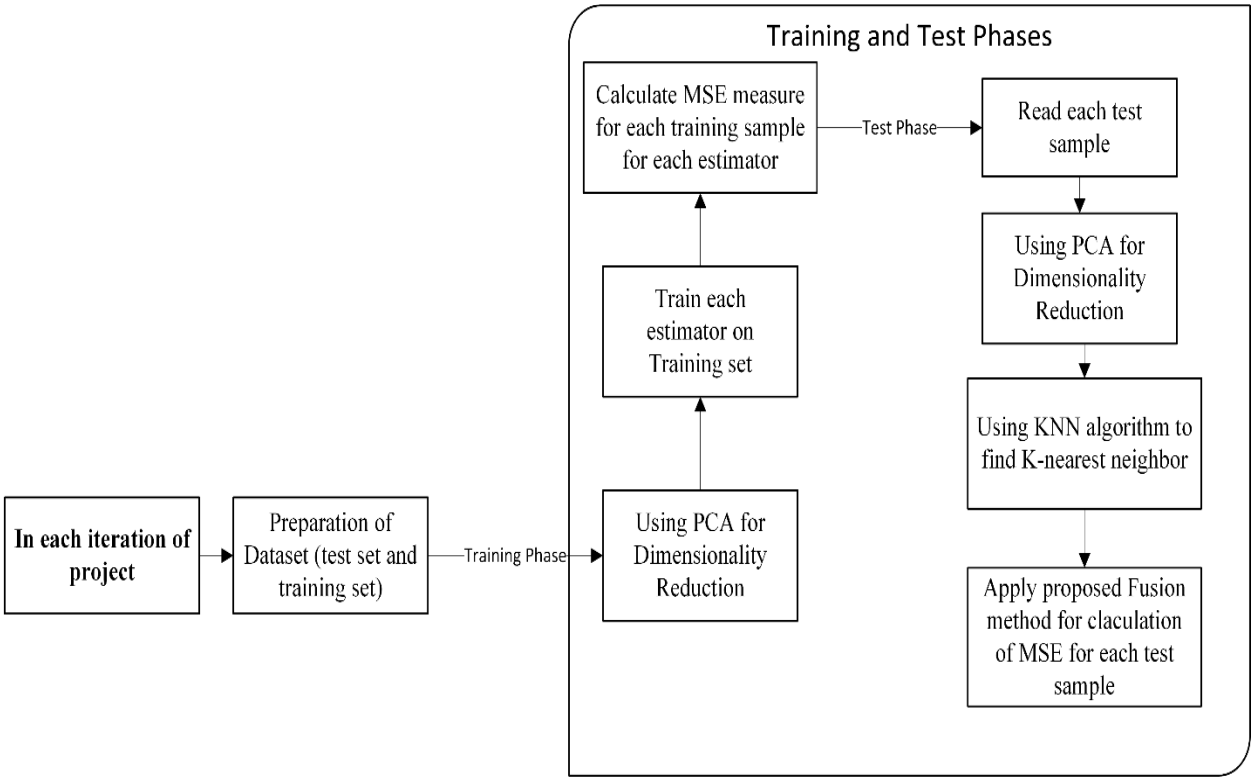


Figure 1. Flowchart of the proposed method

3. Principal Component Analysis

Finding patterns in data with high dimensions is very difficult; therefore, data dimension reduction methods are used to reduce the dimension of data. PCA method is one of the most powerful methods for this purpose [20].With this method, the data dimensions is reduced and also do not ignored a lot of information. Consequently, the estimator can obtain higher accuracy too, by reducing the feature space complexity.

$$COV(X,Y) = \frac{\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y})}{N} \tag{3}$$

Where,COV is the covariance;$\bar{X}$and$\bar{Y}$, are the mean of the variablesX and Y. Also, N is the number of samples. Equation4 is used to calculate the covariance matrix, in which, CM showsthe covariance matrix, and it is a symmetric matrix.

$$CM^{N,N} = (CM(i,j), CM(j,i) = COV(D_i, D_j)) \quad (4)$$

Since, the covariance matrix is a square matrix, eigenvector and eigenvalues are extracted from, covariance matrix. On the other hand, eigenvectors are perpendicular on each other, and length of each is equal to 1.

$$AV = \lambda V \ , \ (A - \lambda I)V = 0 \quad (5)$$

$$\det(A - \lambda I) = 0 \quad (6)$$

In equations 5 and 6, A, is the covariance matrix, and λ, are amount of eigenvalues, and V shows eigenvectors, and I shows the identity matrix. Also det, shows the determinant of a matrix.After calculating the Eigenvalues and eigenvectors, amounts of eigenvalue, indicates the degree of importance, and available information in the eigenvector. Consequently, by ignoring components, with smaller amounts, we can reduce the dimension of the data, appropriately.

4. KNN Algorithm

KNN algorithm, is a kind of methods of Lazy learning and instance-based learning, and is able to find a variety of applications in other areas, such as data processing and image processing [21]. This algorithm relies on the principle of similarity between neighboring samples. In this section, K is a systematic parameter which should be tuned, before. In the KNN method, if the existing sample in database is identified with an equation, and *d*, be size of samples in feature space, then we get:

$$[(x_1, y_1), (x_2, y_2) \dots, (x_N, y_N)] x \in R^d, y \in R \quad (7)$$

Then, by defining a dissimilarity measure, we can find difference, or distance between two samples in the feature space. Different methods have been proposed for dissimilarity measure, that, in short, it can be written as equation 8.

$$D(x_i, x_j)^2 = \|x_i - x_j\|^2 = \sum_{s=1}^{d} (x_{is} - x_{js})^2 \quad (8)$$

$$, x_i = (x_{i1}, x_{i2}, \dots, x_{id})$$

To show the nearest samples in the feature space, we assume the feature space, two-dimensional, and also make the estimation problem into a two-class problem, for simplicity's sake. Nearest neighbors, for samples that were identified, with black color, and placed in the test set, for k = 3, is in Figure 2.

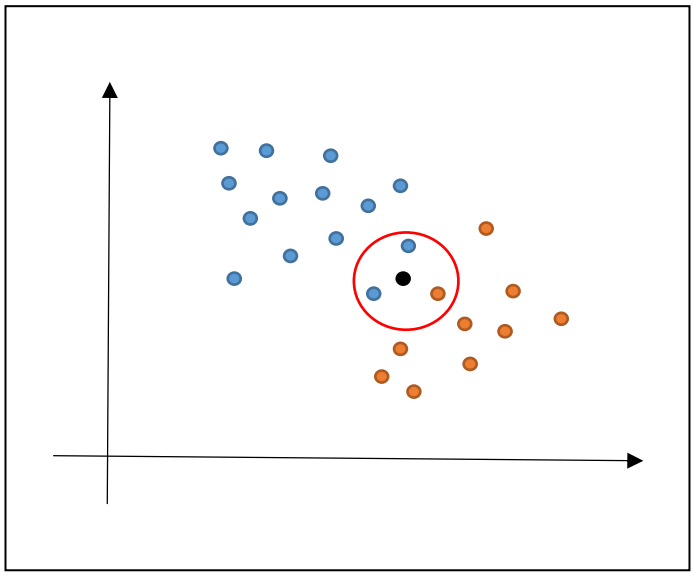


Figure 2. Showing NN-3 in two-dimensional space

5. Experimental Results

In this paper, as stated above, the 3 projects in the article [14], have been used. Available data, in this article have risen with factor 10, in the form of interpolation techniques, and also by the random factor. Three testsets are designed based on three presented data sets, which titled as EX1 and EX2 and EX3.These experiments were designed, fully dynamic, during the progress of the projects. For example, in EX3, it is assumed that the company has done projects I and II, and has started to do the third project. Hence, a data set from first two projects has been formed. Estimation done in beginning of each month, with the progress of the project, and third project information also adds to dataset, dynamically, during its progress.

5.1 Test Number 1

First, Figure 3 shows a cumulative distribution of available information in the connected components for EX3, to determine the number of used components in this project. According to the figure, by using the first three components, we can use almost all available information, in the features, for estimating. As a result, by using 3features rather than 14 features in estimating, the complexity of feature space becomes lower, and thus, the accuracy increases, obviously.

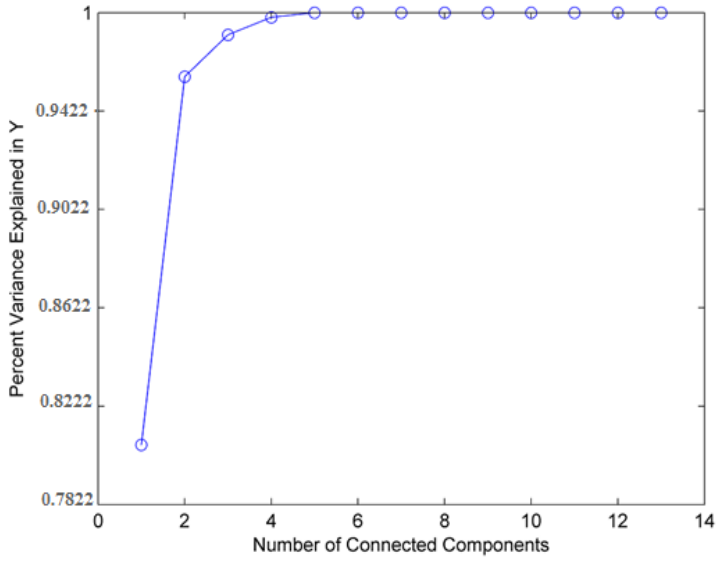


Figure 3. Cumulative distribution graphs, for percentage of useful information for different features

Table 2 shows amount of mean of MSE for different methods. Also dynamic process of development is shown in the figure, where we effort to show, different estimates for different months, in it. In Table 2, the mean values in the different phases of the project are located.

Table 2. Mean of accuracies for the first test for different methods of three different data sets

Number of Neighbor Samples	K = 3				K=5			
Experiments	EX1	EX2	EX3	Average	EX1	EX2	EX3	Average
Proposed Method	0.763	0.501	0.259	0.531	0.766	0.507	0.278	0.540
RT	1.790	1.121	0.557	1.212	1.787	1.121	0.557	1.212
GRNN	0.943	0.931	1.210	1.014	0.943	0.931	1.210	1.014
ETR	1.787	1.434	0.735	1.371	1.790	1.434	0.735	1.371

According to the table, we can see that the accuracy of single model methods is changed extremely, with change in their data sets, so that, neural network is the best estimator in EX1,while in EX3, it is the worst estimator. However, the proposed method, by changing in dataset, coordinates itself, with the best existing local estimator

in ensemble. According to the table, there is relatively little reduction in accuracy of the proposed method, by increasing the number of neighborhoods. According to the figure, the project includes 32 steps, which are divided as follow; project A divided into 12 stages (months), B divided into 11 stages, and C divided into 9 stages for estimation. This division is presented based on the number of divided months, for 3 projects in [14].So that, the first 12 steps, are relating to EX1, and the next 11 steps are related to EX2, and also the last 9 steps are related to EX3.

5.2 Test Number 2

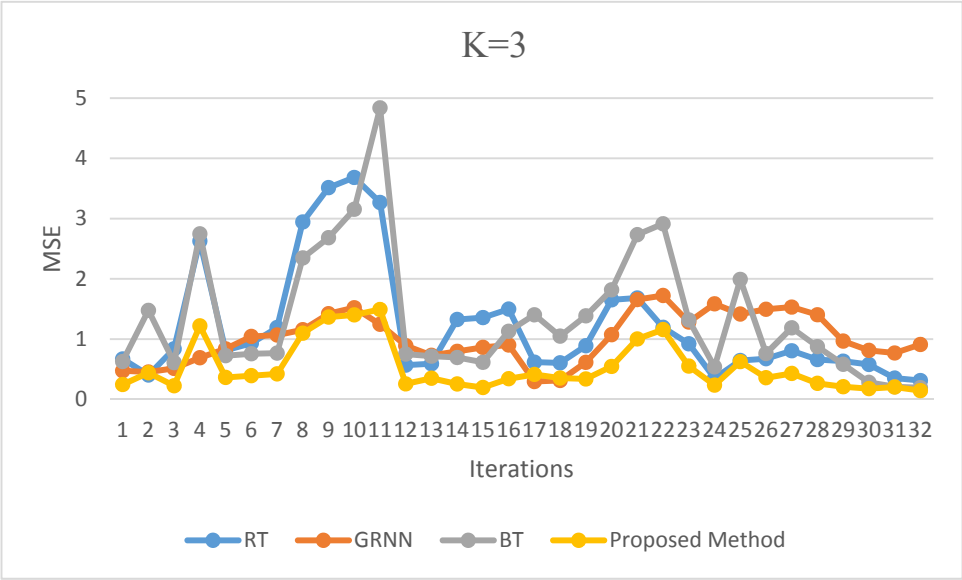
In test number two we attempt to change data, this work, causes to the accuracy of each method in ensemble have change. The results of the second part of experiment are shown in Table 3. This change in this case, is that the method used in [14], in forming dataset, has changed.

Table 3. Mean accuracy of the first experiment for different methods in three different dataset

Number of Neighbor Samples	K = 3				K=5			
Experiments	EX1	EX2	EX3	Average	EX1	EX2	EX3	Average
Proposed Method	0.196	0.266	0.176	0213	0.190	0.264	0.170	0.210
RT	0.438	0.557	0.385	0.464	0.438	0.557	0.385	0.464
GRNN	0.616	1.062	0.780	0.815	0.616	1.062	0.780	0.815
ETR	0.378	0.546	0.341	0.425	0.378	0.546	0.341	0.425

Due to the table, we can see that the accuracy of single model methods could changes dramatically by change their dataset, so that neural network is the best estimator in EX1, while, in EX3, it is the worst estimator. However, proposed method with changing the dataset tune itself with the best existing local estimator in ensemble. Due to the table, in the proposed method, increasing in the number of neighborhoods leads to a little increasing in proposed method accuracy. Also the process of estimation for each month is specified, in the figure 6.

(a)



(b)

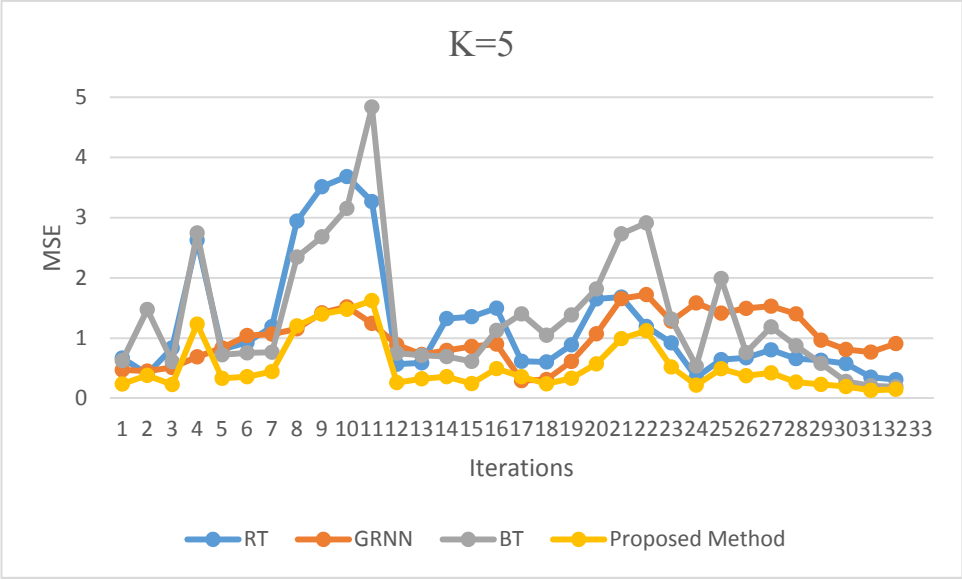
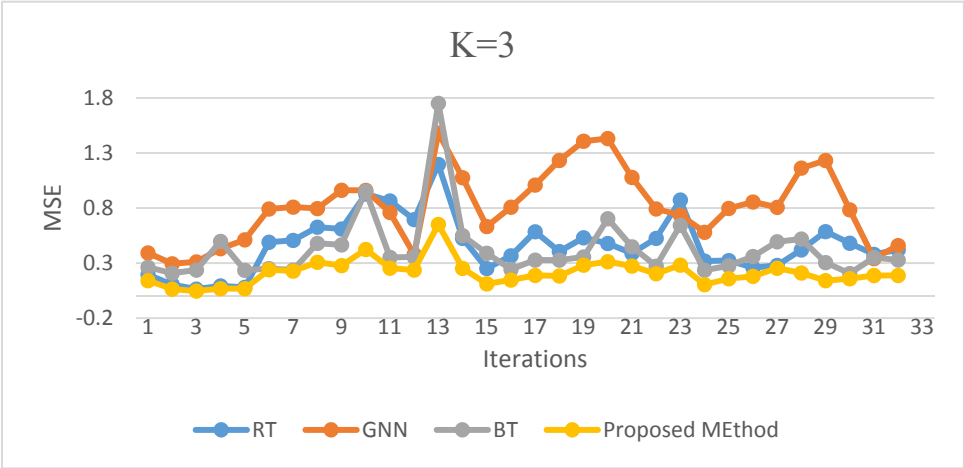


Figure 5. The diagram of four different estimators in the first experiment, at different stages, with different k.

(a)



(b)

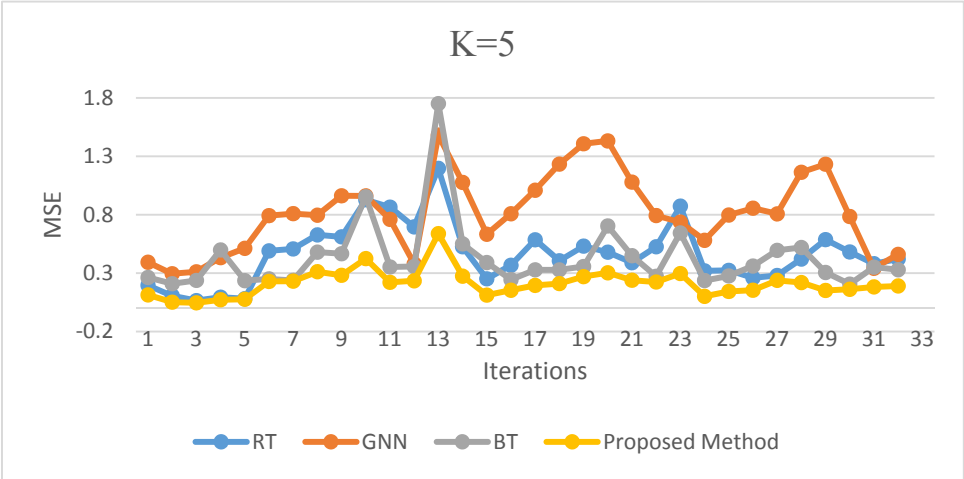


Figure 6. The diagram of four different estimators, in the second experiment, in different stages, with different k.

6. Conclusions

In this paper, a method based on mixture of experts is presented, in which the final result is calculated by the proposed fusion method. KNN method is used to find the best estimators in the neighborhood. Finally, a fusion method is proposed for combination of different estimation. Diversity increases with increasing in deferent learning, and this increase in diversity can increase generalization in final decision. It causes, if the number of estimators, in the test set has increase, or become more complex, not only the accuracy will increase on the training samples, but also this increase in accuracy is obvious in samples testing. As a result, with using more complex learning machines, in ensemble, it can be expected, that the results become far better. The results show that the proposed method has achieved greater accuracy in both tests, compared to each of existing samples in ensemble.

References

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician, 46*(3), 175-185.

Anbari, F. T. (2003). Earned value project management method and extensions. *Project management journal, 34*(4), 12-23.

Barraza, G. A., Back, W. E., & Mata, F. (2000). Probabilistic monitoring of project performance using SS-curves. *Journal of Construction Engineering and Management, 126*(2), 142-148. http://dx.doi.org/10.1061/(ASCE)0733-9364(2000)126:2(142)

Chao, L. C., & Chien, C. F. (2009). Estimating project S-curves using polynomial function and neural networks. *Journal of Construction Engineering and Management, 135*(3), 169-177. http://dx.doi.org/10.1061/(ASCE)0733-9364(2009)135:3(169)

Cheng, M. Y., Hoang, N. D., Roy, A. F., & Wu, Y. W. (2012). A novel time-depended evolutionary fuzzy SVM inference model for estimating construction project at completion. *Engineering Applications of Artificial Intelligence, 25*(4), 744-752. http://dx.doi.org/10.1016/j.engappai.2011.09.022

Cheng, M. Y., Peng, H. S., Wu, Y. W., & Chen, T. L. (2010). Estimate at completion for construction projects using evolutionary support vector machine inference model. *Automation in Construction, 19*(5), 619-629. http://dx.doi.org/10.1016/j.autcon.2010.02.008

Duda, R. O., Hart, P. E., & David, G. S. (2012). *Pattern classification*. John Wiley & Sons.

Harandi, M. T., Ahmadabadi, M. N., Araabi, B. N., Bigdeli, A., & Lovell, B. C. (2010). Directed Random Subspace Method for Face Recognition, 20th International Conference on Pattern Recognition (ICPR), 2688-2691.

Howes, & Rodney (2000). Improving the performance of Earned Value analysis as a construction project management tool. *Engineering Construction and Architectural Management, 7*(4), 399-411. http://dx.doi.org/10.1046/j.1365-232X.2000.00171.x

Hsieh, N. C., & Hung, L. P. (2010). A data driven ensemble classifier for credit scoring analysis. *Expert Systems with Applications, 37*(1), 534-535. http://dx.doi.org/10.1016/j.eswa.2009.05.059

Iranmanesh, S. H., & Zarezadeh, M. (2008). Application of Artificial Neural Network to Forecast Actual Cost of a Project to Improve Earned Value Management System. *Proceeding of World Academy of Science, engineering and technology, 32*, 2070-3740.

Jacob, D. (2003). Forecasting project schedule completion with earned value metrics. *The Measurable News, 1*, 7-9.

Khandare, M., A., & Vyas, G. S. (2012). Project Duration Forecasting Using Earned Value Method and Time Series. *International Journal of Engineering and Innovative Technology (IJEIT), 1*(4).

Li, K., & Wang, L. X. (2009). Ensemble Methods of Face Recognition Based on Bit-plane Decomposition. *International Conference on Computational Intelligence and Natural Computing, 1*, 194-197. http://dx.doi.org/10.1109/CINC.2009.216

Lipke, & Walt. (2003). Schedule is different. *The Measurable News*, *31*(4).

Mohammad, T., & Hajiali, K. S. (2014). The Dynamic Model of Estimating the Time and Cost of Project Completion in an Environment of Uncertainty. *Emergencias Engineering and Technology Issue, 2*(2).

Shahanaghi, K., & Hajiali, M. T. (2014). Estimation of Project Time and Cost at Completion Using Fuzzy Kalman Filter and ARMA Model, Emergencias Special Management. *Business and Economics, 2*(1).

Son, H., Kim, C., & Kim, C. (2012). Hybrid principal component analysis and support vector machine model for predicting the cost performance of commercial building projects using pre-project planning variables. *Automation in Construction, 27*, 60-66. http://dx.doi.org/10.1016/j.autcon.2012.05.013

Vandevoorde, S., & Vanhoucke, M. (2006). A comparison of different project duration forecasting methods using earned value metrics. *International journal of project management, 24*(4), 289-302. http://dx.doi.org/10.1016/j.ijproman.2005.10.004

Varol, E., Gaonkar, B., Erus, G., Schultz, R., & Davatzikos, C. (2012). Feature ranking based nested support vector machine ensemble for medical image classification, 9th IEEE. *International Symposium on Biomedical Imaging (ISBI),* 146-149.

Zikeba, M., & Wikatek, J. (2012). Ensemble classifier for solving credit scoring problems. *Technological Innovation for Value Creation,* 59-66.