Bulletin of Electrical Engineering and Informatics
Vol. 10, No. 5, October 2021, pp. 2789~2795
ISSN: 2302-9285, DOI: 10.11591/eei.v10i5.2785

A novel imbalanced data classification approach using both under and over sampling

Seyyed Mohammad Javadi Moghaddam[1], Asadollah Noroozi[2]
[1]Department of Computer Engineering, Bozorgmehr University of Qaenat, Qaen, Iran
[2]Department of Civil Engineering, California, USA

Article Info

Article history:

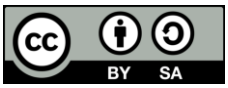
Received Jan 6, 2021
Revised Apr 13, 2021
Accepted Aug 2, 2021

Keywords:

Cluster-based approach
Imbalanced data
Over-sampling
SMOTE algorithm
Under-sampling

ABSTRACT

The performance of the data classification has encountered a problem when the data distribution is imbalanced. This fact results in the classifiers tend to the majority class which has the most of the instances. One of the popular approaches is to balance the dataset using over and under sampling methods. This paper presents a novel pre-processing technique that performs both over and under sampling algorithms for an imbalanced dataset. The proposed method uses the SMOTE algorithm to increase the minority class. Moreover, a cluster-based approach is performed to decrease the majority class which takes into consideration the new size of the minority class. The experimental results on 10 imbalanced datasets show the suggested algorithm has better performance in comparison to previous approaches.



Corresponding Author:

Seyyed Mohammad Javadi Moghaddam
Department of Computer Engineering
Bozorgmehr University of Qaenat, Qaen, Iran
Email: smjavadim@buqaen.ac.ir

1. INTRODUCTION

An essential challenge faced by the traditional classification algorithms is the distribution of data where the classes are imbalanced. For example, healthy transactions are significantly bigger than fraudulent transactions. In this situation, the classifiers trend to the majority class and ignore the minority one. There are three categories for classical imbalanced data classification approaches. The algorithmic level methods that try to strengthen the classification algorithm to enforce the learning towards the minority samples [1], [2]. The second group of approaches is ensembles classifiers that contain two methods [3], [4]: bagging and boosting. Bagging includes various classifiers that are used to subsets of the dataset [5]. Likewise, in boosting, the complete dataset is applied to train classifiers so that it gives more attention to the samples that are misclassified [6], [7]. The third category entails scenarios such as pre-processing the data to balance before providing as the input data or improving the classifiers. The data processing known data level techniques are preferred as it has vast applications [8], [9].

The major objective of data level algorithms is to either decreasing or increasing the class number. These approaches try to achieve the same sample number for both classes. The under-sampling approach attempts to reduce the instances of the majority class. This technique discards useful information which could be essential for classifiers. Moreover, it is an inaccurate representation of the population. The over-sampling algorithm increases the minority class number by replicating samples [10], [11]. Unlike under-sampling, this approach leads to no information loss. However, it increases the probability of overfitting because of reproducing the minority class samples [12], [13].

To overcome the challenges of under and over sampling algorithms, some researchers proposed to combine the approach with other techniques. Khan *et al*. [14] described an approach in which a cost-sensitive method based on the neural networks can train representations of the feature for both the minority one and the majority category. This method tries to improve the classifier. Therefore, the original data is no change. Castellanos *et al*. [15] suggested a strategy based on a string converting. This strategy converts the SMOTE technique to a string space. The improvement of this method is 97.5 according to the F-measure score.

Many researchers have tried to perform a clustering method to balance the classes. Prachuabsupakij *et al*. [16] suggested an approach in which a k-means based method decreases the overlapping of the classes. This method clusters the original dataset into two classes. Then, a clustering switching method and the SMOTE technique are performed on each class. The output is two balanced training set. This model clusters the majority class into two classes without regarding the size of the minority class. The average F-measure of this model was 0.975. Czarnowski *et al*. [17] presented an approach based on clustering where similarity coefficient computed for samples of each data class independently. Then, similar samples are clustered into the same class. The maximum accuracy of this method is 98.01. Lin *et al*. [18] proposed an under-sampling method using the clustering technique that the majority data is divided into k class. This algorithm calculates k regarding the minority class size. The best average classification accuracy is 0.904.

This paper introduces a combinatorial algorithm to overcome the imbalanced problem. It tries to produce the minority class item by the SMOTE method. Likewise, it uses a clustering algorithm to decrease the majority class. Unlike previous approaches, it clusters the majority one regarding the new minority one. The novelty of this work is that the rate of increase of the minority class and decrease of the corresponding majority class is done together. The paper has been arranged as; the next section includes some basic techniques relevant background and the proposed algorithm, section 3 and 4 provide the results of the experiments. Finally, concluding points are in section 5.

2. RESEARCH METHOD

Before the proposed algorithm was introduced, a summary of the basic knowledge would be presented. This work uses SMOTE technique to increase the minority classes number. Moreover, the approach uses a clustering method as an under-sampling algorithm to decrease the majority class.

2.1. SMOTE technique

This algorithm carries out an over-sampling method to balance the imbalanced data [19]. The major idea of the method is to produce synthetic samples. The new instance is created according to the interpolation of some samples in minority class that are neighborhood space. Therefore, it focuses on the feature aspect instead of the data one. In other words, the method considers both the value of features and the relationship between them [12]. Figure 1 depicts a simple example of SMOTE. First, a minority class sample χ_i is considered to produce a new synthetic point. Then, several nearest neighbors regarding a distance metric are selected. Finally, k samples are selected in a random way to obtain the new samples by insertion (χ_i to χ_k). Therefore, the distance between the considered instance and its neighbors is multiplied by a random coefficient between 0 and 1. Consequently, some new points are added which one is chosen at random (rd_1 to rd_k).

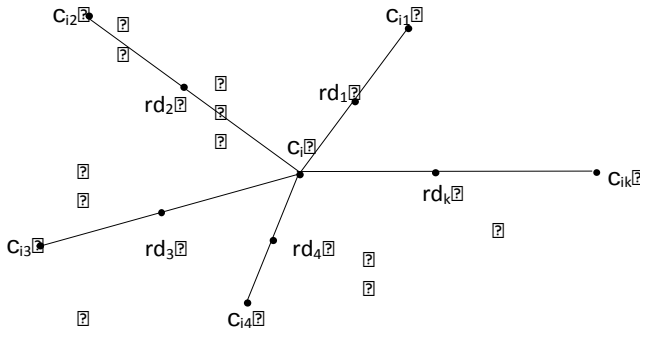


Figure 1. Producing the new SMOTE point

2.2. K-means algorithm

The k-means method is widely used in the machine learning area. It is an iterative technique that attempts to divide the dataset into k distinct cluster where each data item belongs to only one cluster [20].

Moreover, if the distances of the data points and each centroid of the cluster are calculated, the sum of them should be at the minimum. In other words, there is less variation within the clusters. The phases of the algorithm include:

- Compute the clusters number (K)
- Select k centroids from data points randomly
- Keep iterating until the centroids are constant
- Calculate the sum of the distance of data points and the centroids
- Determine the centroids of each cluster regarding the average of the data point of the cluster

$$J = \sum_{i=1}^{m} \sum_{j=1}^{K} w_{ij} ||x^i - \mu_j||^2 \quad (1)$$

2.3. Proposed algorithm

This paper tries to combine both under and over sampling approaches. The proposed method performs the SMOTE algorithm on the minority class to increase its samples. Moreover, it uses a clustering method to decrease the majority class without losing data. The phases of the algorithm are as:

- Performing SMOTE technique on minority cluster
- Computing the clusters number by proportion the majority size and the size of the new minority one
- Performing the k-means algorithm on majority cluster
- Combine each cluster with new minority class
- Performing a classifier for each class
- Classification with maximum probability vote

Figure 2 shows the flow chart of proposed algorithm. The method tries to increase the minority class by considering the IR of the dataset. The number of clusters (known K) is determined according to the new minority size and majority size. The value of K is equal to the size of the majority class divided by the size of the minority class. In the next step, the K-mean algorithm is performed on the majority class to produce k clusters. Then, each cluster is combined with the new minority class. A classifier categorizes each new cluster. Finally, the model selects the cluster with maximum probability vote.

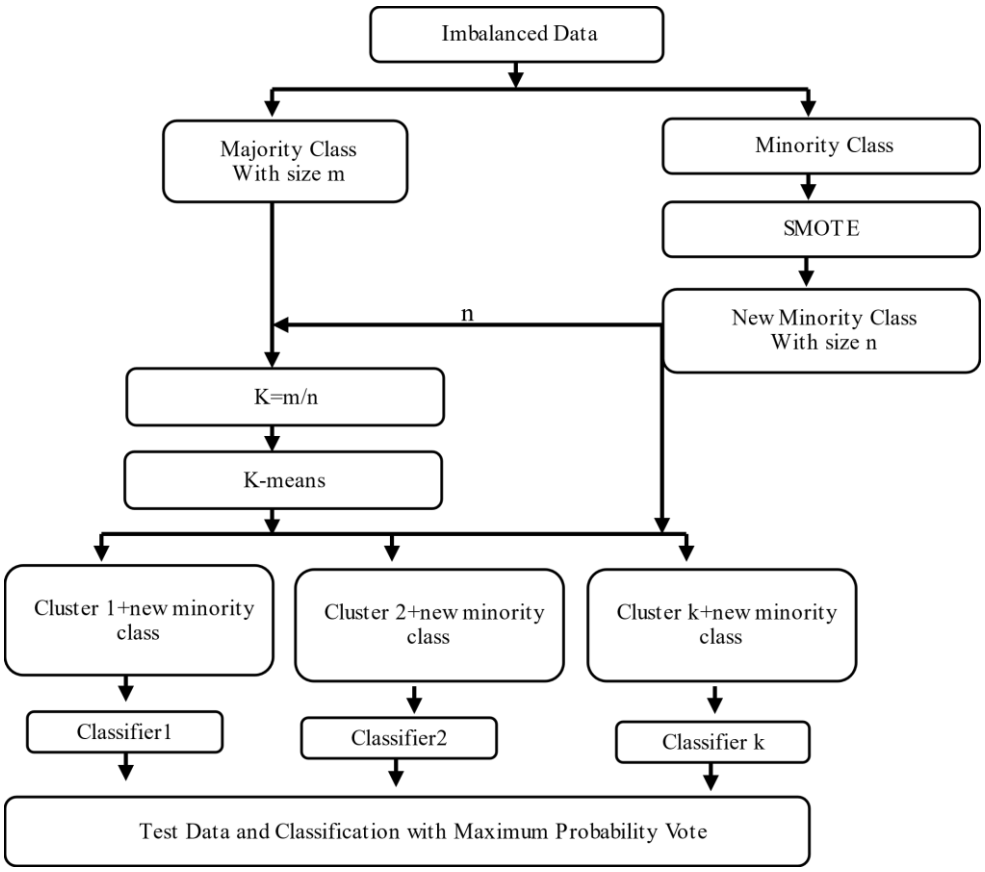


Figure 2. Flow chart of the proposed algorithm

3. RESULTS AND DISCUSSION

3.1. Dataset and experimental setting

The experimental datasets are all from the KEEL repository [21]. The datasets have a various imbalanced ratio. The number of data samples is from 214 to 5472. Table 1 shows the experimental parameters of them.

Table 1. Datasets information

No.	Dataset	Attributes	Number of Samples	IR
1	glass1	9	214	1.82
2	pima	8	768	1.87
3	vehicle3	18	846	2.99
4	ecoli1	7	336	3.36
5	new-thyroid1	5	215	5.14
6	ecoli2	7	336	5.46
7	page-blocks	10	5472	8.79
8	yeast6	8	1484	41.4
9	poker-8-vs-6	10	1477	85.88
10	abalone19	8	4174	129.44

This paper performs SMOTE algorithms to increase minority instances. The number of oversampling instances is determined according to the IR of the dataset. Then, the number of clusters (k) is calculated regarding the new minority cluster and the majority size. In the next step, the K-means method produces the clusters. Then, each cluster is combined with the new minority class. Classification is done for each new dataset. Finally, voting selects the best one. Figure 3 presents the proposed algorithm steps in detail. To evaluate the classification by the proposed algorithm, four different classifiers were performed including decision tree [22], support vector machine (SVM) [23], nearest neighbor classifiers [24], and ensemble classifiers [19].

The proposed Algorithm
Input:
1) Given $\{(x_1,y_1),...,(x_n,y_n)\}$ $x_i \in X$, $y \in Y=2$
2) M_i= the minority cluster number
3) M_j= the majority cluster number
4) $IR=M_j/M_i$
5) K is the clusters number

6) $New\_M_i = SMOTE(M_i)$
7) $K=M_j / New\_M_i$
8) $C_{1..k} = Kmeans(M_j, k)$
9) **For** dc=1 to K do
10) {
11) $SC_i = New\_M_i \cup C_d$
12) $V_i = baseclassifier(SC_i)$
13) }
Output: $V^* = voting\ V_i$ regarding maximum probability

Figure 3. The proposed algorithm pseudo code

3.2. Evaluation methods

Accuracy calculates the correct predicted instances number over the all instances

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

where: FP is an outcome that indicates something is present when really is not
FN is a result that presents negative when it should not
TP is an upshot indicates positive when really is
TN is a result that shows negative when really is

But if the distribution is unbalanced it can be misleading. If the distribution is unbalanced accuracy can be misleading. Therefore, it is better to rely on precision and recall. Likewise, in the same way, a Precision-Recall curve is suitable to evaluate the classifier in an imbalanced class. Moreover, the region under the curve known AUC is a performance measurement for the classification method.

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

Table 2 presents the accuracy of the approaches on 10 datasets. Moreover, Table 3 shows the AUC for all datasets in the three situations. For comparing the performance of the suggested approach, the classification results for the normal dataset, dataset after performing SMOTE, a density-based under-sampling algorithm (DBU) [25] and the proposed method outcome were evaluated. As mentioned before, it is easy to get a high accuracy without actually making a suitable prediction when there are imbalanced classes. Therefore, precision and recall were computed. Then, the area under the precision-recall curves (AUC) was used as a summary of the model performance. A precision-recall curve shows a balance achieved between the TP rate and the positive value that the model predicts using different probability thresholds. Table 3 presents the AUC for all datasets in the three situations.

Table 2. Accuracy for all datasets

No.	Dataset	Normal %	SMOTE %	The proposed Metod %
1	glass1	84.3	93.4	98.9
2	pima	77.5	83.9	83.9
3	vehicle3	85.2	88.7	93
4	ecoli1	90.2	92.5	99.3
5	new-thyroid1	99.1	99.2	99.6
6	ecoli2	96.1	97.2	97.8
7	page-blocks	89.3	98	99.3
8	yeast6	98.4	96.3	97.7
9	poker-8-vs-6	99.6	99.7	99.8
10	abalone19	99.2	99.1	99.6

Table 3. AUC for all datasets

No.	Dataset	Normal	SMOTE	DBU	The proposed Method
1	glass1	0.916095	0.989536	0.954325	0.994768
2	pima	0.830507	0.912511	0.736557	0.912511
3	vehicle3	0.899753	0.943608	0.758543	0.976075
4	ecoli1	0.972547	0.973063	0.797543	0.989592
5	new-thyroid1	0.982937	0.999921	0.985654	0.999996
6	ecoli2	0.946167	0.990605	0.984432	0.992727
7	page-blocks	0.927108	0.998655	0.943543	0.998866
8	yeast6	0.968653	0.969167	0.859876	0.978043
9	poker-8-vs-6	0.881668	0.969903	0.943754	0.998708
10	abalone19	0.608115	0.923286	0.652543	0.990296

The proposed algorithm combines oversampling and under-sampling techniques. The decreasing rate of majority class is done by considering the rate of increasing minority class. Therefore, the proposed method uses more instances of the original data. Moreover, the increasing rate of minority class is according to the IR of the dataset. The results show increase accuracy and AUC of the proposed model compared to the SMOTE method on benchmark imbalanced datasets from the KEEL repository.

4. CONCLUSION

This work proposes a novel technique to bias imbalanced data. To overcome the unbalance problem, both under and over-sampling approaches are used. Most imbalanced data classification techniques try to balance the data using increasing the minority class or decreasing the majority class that results in changing the original data. The proposed algorithm tries to reduce the changing rate of the primary dataset. The algorithm firstly performs the SMOTE method on the minority cluster to produce new instances. Then, the k-means clustering algorithm decreases the majority class, which considers k regarding the size of the new minority class. Finally, each cluster and the new minority class are considered as the input data of a classifier.

The experimental results show increase accuracy and AUC of the proposed model compared to the SMOTE method on benchmark imbalanced datasets from the KEEL repository. The method is further performed on other real-world engineering datasets.

ACKNOWLEDGEMENTS

The author would like to acknowledge the financial support of the Bozorgmehr University of Qaenat for this research under contract number 39141.

REFERENCES

[1] S. Saryazdi, B. Nikpour, and H. Nezamabadi-Pour, "NPC: Neighbors' progressive competition algorithm for classification of imbalanced data sets," *3rd Iranian Conference on Intelligent Systems and Signal Processing ICSPIS*, 2017, pp. 28-33, doi: 10.1109/ICSPIS.2017.8311584.

[2] S. Saeed and H. C. Ong, "A bi-objective hybrid algorithm for the classification of imbalanced noisy and borderline data sets," *Pattern Analysis and Applications*, pp. 1-20, 2018.

[3] B. Wang, J. Pineau, "Online Bagging and Boosting for Imbalanced Data Streams," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 12, pp. 3353-3366, 1 Dec 2016, doi: 10.1109/TKDE.2016.2609424.

[4] T. T. Khuat and M. H. Le, "Evaluation of sampling-based ensembles of classifiers on imbalanced data for software defect prediction problems," *SN Computer Science*, vol. 1, no. 2, pp. 1-16, 2020, doi: 10.1007/s42979-020-0119-4.

[5] K. Napierala and J. Stefanowski, "Modifications of classification strategies in rule set based bagging for imbalanced data," in *International Conference on Hybrid Artificial Intelligence Systems*, Springer, pp. 514-525, 2012.

[6] J. Du, C.-M. Vong, C.-M. Pun, P.-K. Wong, and W.-F. Ip, "Post-boosting of classification boundary for imbalanced data using geometric mean," *Neural Networks*, vol. 96, pp. 101-114, 2017, doi: https://doi.org/10.1016/j.neunet.2017.09.004.

[7] Q. Li and Y. Mao, "A review of boosting methods for imbalanced data classification," *Pattern Analysis and Applications*, vol. 17, no. 4, pp. 679-693, 2014.

[8] S. M. Augusty and S. Izudheen, "A survey: evaluation of ensemble classifiers and data level methods to deal with imbalanced data problem in protein-protein interactions," *Review of Bioinformatics and Biometrics*, vol. 2, no. 1, pp. 1-9, 2013.

[9] R. A. Hamad, M. Kimura, and J. Lundström, "Efficacy of imbalanced data handling methods on deep learning for smart homes environments," *SN Computer Science*, vol. 1, no. 4, pp. 1-10, 2020, doi: 10.1007/s42979-020-00211-1.

[10] M. Koziarski, B. Krawczyk, and M. Woźniak, "Radial-based approach to imbalanced data oversampling," in *International Conference on Hybrid Artificial Intelligence Systems*, Springer, pp. 318-327, 2017, doi: 10.1007/978-3-319-59650-1_27.

[11] S. Bej, N. Davtyan, M. Wolfien, M. Nassar, and O. Wolkenhauer, "LoRAS: An oversampling approach for imbalanced datasets," *Machine Learning*, vol. 110, no. 2, pp. 279-301, 2021, doi: 10.1007/s10994-020-05913-4.

[12] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," *Journal of Artificial Intelligence Research*, vol. 61, pp. 863-905, 2018, doi: 10.1613/jair.1.11192.

[13] B. Krawczyk, C. Bellinger, R. Corizzo, and N. Japkowicz, "Undersampling with Support Vectors for Multi-Class Imbalanced Data Classification," *International Joint Conference on Neural Networks (IJCNN 2021)*, July 2021.

[14] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel and R. Togneri, "Cost-Sensitive Learning of Deep Feature Representations From Imbalanced Data," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 8, pp. 3573-3587, Aug. 2018, doi: 10.1109/TNNLS.2017.2732482.

[15] F. J. Castellanos, J. J. Valero-Mas, J. Calvo-Zaragoza, and J. R. Rico-Juan, "Oversampling imbalanced data in the string space," *Pattern Recognition Letters*, vol. 103, pp. 32-38, 2018.

[16] W. Prachuabsupakij and S. Simcharoen, "A Cluster Switching Method for Sampling Imbalanced Data," in *Proceedings of the 2nd International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, ACM, pp. 12-16, 2018.

[17] I. Czarnowski and P. Jędrzejowicz, "Cluster-Based Instance Selection for the Imbalanced Data Classification," in *International Conference on Computational Collective Intelligence*, Springer, pp. 191-200, 2018.

[18] W.-C. Lin, C.-F. Tsai, Y.-H. Hu, and J.-S. Jhang, "Clustering-based undersampling in class-imbalanced data," *Information Sciences*, vol. 409, pp. 17-26, 2017, doi: 10.1016/j.ins.2017.05.008.

[19] Q. Wang, Z. Luo, J. Huang, Y. Feng, and Z. Liu, "A novel ensemble method for imbalanced data learning: bagging of extrapolation-SMOTE SVM," *Computational intelligence and neuroscience*, vol. 2017, ID. 1827016, pp. 1-11, 2017, doi: 10.1155/2017/1827016.

[20] Y. Lu, Y. -M. Cheung, Y. Y. Tang, "Self-Adaptive Multiprototype-Based Competitive Learning Approach: A k-Means-Type Algorithm for Imbalanced Data Clustering," *IEEE Transactions on Cybernetics*, vol. 51, no. 3, pp. 1598-1612, March 2021, doi: 10.1109/TCYB.2019.2916196.

[21] J. Alcalá-Fdez *et al.*, "Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework," *Journal of Multiple-Valued Logic & Soft Computing*, vol. 17, pp. 255-287, 2011.
[22] J.-S. Lee, "AUC4. 5: AUC-based C4. 5 decision tree algorithm for imbalanced data classification," *IEEE Access*, vol. 7, pp. 106034-106042, 2019, doi: 10.1109/ACCESS.2019.2931865.
[23] P. Liang, F. Zheng, W. Li, and J. Hu, "Quasi- linear SVM classifier with segmented local offsets for imbalanced data classification," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 14, no. 2, pp. 289-296, 2019, doi: 10.1002/tee.22808.
[24] S. Susan and A. Kumar, "Learning Data Space Transformation Matrix from Pruned Imbalanced Datasets for Nearest Neighbor Classification," *IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems*, 2019, pp. 2831-2838, doi: 10.1109/HPCC/SmartCity/DSS.2019.00397.
[25] Y. Hou, B. Li, L. Li, and J. Liu, "A density-based under-sampling algorithm for imbalance classification," *Journal of Physics: Conference Series*, 2019, vol. 1302, no. 2, p. 1-11, doi: 10.1088/1742-6596/1302/2/022064.

BIOGRAPHIES OF AUTHORS

Seyyed-Mohammad Javadi-Moghaddam was born in Qaen, Iran, in 1975. He received the B.E. degree in computer engineering from the Ferdowsi University of Mashhad, Iran, in 1998, the MSc degree in Computer Software from AZAD University of Mashhad, Iran, in 2007, and Ph.D. degree in Computer Science from National Tech- nical University of Athens, Greece in 2016. In 2007, he joined the Department of Computer Engineering, PNU University of Iran, as a Lecturer. Since September 2016, he is with the Department of Computer Engineering, Bozorg- mahr University of Qaenat, Iran as an Assistant Professor. His current research interests include Cloud computing, Imbalanced data, and distributed systems.

Asadollah Noroozi was born in Qaen, Iran, in 1974. He received the B.E. degree in civil engineering from the Ferdowsi University of Mashhad, Iran, in 1996, the MSc degree in road engineering from Amirkabir University of Tehran, Iran, in 1999. He is a project engineering in California department of transportation since 2016.