JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A New Framework for Finding Optimal k Value in k-fold Cross Validation Technique in Machine Learning Applications

Dr. D. Mabuni
Assistant Professor
Dept. of Computer Science and Technology
Dravidian University, Kuppam, Andhra Pradesh, India

***Abstract*:** The impact of artificial intelligence and machine learning is very high on the society and in particular Machine Learning applications usages are increasing rapidly everyday and they became indispensible with the daily activities of human beings. Accurate learning models are very much useful nowadays and in particular decision tree classification models are needed in diversified domain applications. The popular k-fold cross-validation technique is particularly useful in determining the predictive performance of the model. Previous research communities had already decided that k value in k-fold cross validation must always be 10 or 20 but in real time situations it is not always true. That is sometimes k value may be 5 or 6 or 15 and so on. In the present study by employing many standard and UCI machine learning datasets experiments are conducted. After inspecting and observing experimental results thoroughly new postulates such as k value in k-fold cross validation may be taken into consideration other than usual k = 10 or k = 20 in some applications. That is, always k value should not be 10 or 20 instead k may take on any positive integer (k > 1) value.

Index Terms: **optimal k value, k-fold cross validation, k = 10 or k = 20, predictive performance, decision tree classification, machine learning models.**

I. INTRODUCTION

Cross-validation in machine learning is very popular and one of the most widely used and selected data re-sampling techniques in order to evaluate the generalization ability and capabilities of a known predictive data model and to prevent usually occurring over-fitting problems. To build the best model for the prediction of real future cases, the learning algorithm is generally applied to the entire learning dataset. This model should not be cross-validated. The actual purpose of cross-validation in the model building phase is to provide an approximate estimate for the performance of this original model on new data. In general, feature selection is considered as an integral part of the actual model building process. The intention of cross-validation technique is to find an approximate estimator of the final model on the new data. There are no clearly stated standards, procedures, algorithms and rules for selecting the best value of k for k-fold cross-validation and in many real and scientific experiments there is a need and lot of demand to select an intelligent, state-of-the-art and correct k value. In particular, cross-validation technique is specifically useful for fine tuning the model parameters in order to estimate true prediction errors of models.

Cross validation is the most popular data sampling method used to evaluate machine learning models when the dataset size is small. It uses one parameter called k that represents groups or folds of the single datasets. Cross validation is generally used for estimating the real skill of the machine learning model on the unseen data, which is not used in any part of the training and validation datasets. The results of k-fold cross validation are usually presented in the form of statistical parameters such as mean, variance and standard deviation. In the machine learning literature there exist many variations of cross validation technique. Cross validation technique is frequently used to validate the efficiency of the model by taking training and testing sub sets of the given dataset. In machine learning it is always a good practice of testing and knowing the stability of the model. Some techniques for finding the stability of the model are given in the following list:

1) Stratified Cross-Validation.
2) Nested or double cross validation
3) Repeated

4) LOOCV
5) Train/Test Split

Each method has got its merits and demerits

1) In stratified cross validation the size of each fold is same and at the same time all the class proportions in each fold is same.
2) In the nested cross-validation a separate cross-validation is applied within each fold of the main cross validation fold.
3) In repeated cross validation the cross validation technique is applied repeatedly n number of times in each main fold.
4) LOOCV the abbreviation is leave-one-out cross validation. It is an extreme case where one instance is used as a test instance and all the remaining instances are used as training instance dataset.
5) Train/Test split. In this technique a model is evaluated by dividing the given set into training and testing datasets.

In general, a model is trained using the selected training dataset and the same model is tested using validation dataset and once the model is validated then finally it is tested using test dataset which is not used in both training and validation. In the case of k-fold cross validation same experiment is repeated k times and the final parameters are computed by summarizing the results. For example, in the case of decision tree classification accuracy, average of all the validation accuracy results is computed.

Finding optimal k value in k-fold cross validation method may require many parameters including number of folds and number of iteration to be considered. Cross validation is the most popularly used data splitting method for model validation and consequently for model selection. In cases when k is small bias is more and in cases when k is large variance is more and the correct value of the k must be balanced between these two special parameters.

The main goal of AdaBoost is to combine two or more selected weak learners for the purpose of improving classifier performance. AdaBoost iteratively trains the selected weak learners and find an optimal or robust weight for each learner separately. Final classifier is computed by combining all the weighted individual learners.

II. RELATED WORK

In many of the real time applications one training and one test dataset is not sufficient for obtaining accurate and realistic classification accuracy results. Definitely there is a need of applying many training and testing pairs of datasets for getting reasonably correct and accurate classification results. As a result of this datasets must be divided appropriately into groups of training, validation and test sets. Cross validation technique and in particular k-fold cross validation technique is one such very important machine learning technique for finding the optimal classification accuracy results in all those cases where training datasets are of small sizes by their inherent nature.

For larger training datasets there will not be a problem of dividing the dataset into more number of training and validation pairs of sets for experimental setup purpose but there is a problem with small training datasets during the division process of many pairs of training and validation sets.

There exist many common tactics for selecting the best k value for a given training dataset. The general rule is that fit and evaluate any machine learning model on the training dataset and then finally verify the model prediction accuracy on the held-back test dataset. That is a separate test dataset must be used for testing the fine tuned machine models. In many real cases extensive, sensitive analysis ideas are required for finding k value and some of the sensitive ideas are:

1) Model behavior,
2) Model correctness,
3) Model performance,
4) Model applicability,
5) Model capacity and so on.

Numerous data classification techniques are available in machine learning. Kaushika Pal and Biraj. V. Patel [1] experimentally tried to find the best classification algorithm for classifying the documents. The experimental results of various classification algorithms are computed and then verified with holdout accuracy estimator and the k-fold cross validation accuracy estimator for reliability checking. Authors experimentally proved that support vector machines, naïve Bayesian, and random forests classification algorithms are far better than the decision trees and k-nearest neighbor classification algorithms.

A fold in a cross validation consists of one training dataset and one test dataset. In k-fold cross validation accuracy is estimated for each fold separately and then finally over all accuracy is estimated by aggregating the accuracies of all these folds. Here the main important and also the fundamental point is that there are many possible ways for dividing the given dataset into k-folds such that in each fold one training and one validation set is present. Dividing the dataset into folds is really considered as an optimization problem. Gunasegaran and Yu-N Cheah [2] proposed an evolutionary cross validation algorithm for finding optimal folds of the given dataset for improving

predictive modeling accuracy. Experimental results have shown that the results of the proposed algorithm are superior to the k-fold cross validation method.

Tamilarasi and Uma Rani [3], have used popular six different types of machine learning algorithms such as decision trees, k-nearest neighbor (KNN), Naïve Bayes technique, Classification and Regression Tree (CART), Linear Regression and support vector machines (SVM) to find similar characteristics on crime data and then these algorithms are tested for finding accuracy of the data. The main goal of this research is to find efficiency and real world applicability of these classification algorithms.

Prediction accuracy is the most important technique in many data modeling techniques and there is no exception in modeling the ground water flow. Nurhayati et al. [4] have applied and used extreme learning machine technique for modeling the ground water flow tidal lowland reclamation. The model accuracy is measured using hold-out and k-fold cross validation methods. The experimental results have shown that k-fold cross validation accuracy is very much better than hold-out accuracy.

The fundamental task of molecular diagnostics is to find a set of biomarkers for accurate prediction of genetic details in the case of complex diseases of patients. In such cases the classification model must be compact, high accurate and more generalizable. Venkatesh et al. [5] proposed an iterative technique that is effective, efficient, and applicable to many datasets particularly for genomic datasets. This technique is particularly useful when independent dataset is not available separately. The proposed technique is practically applied on many real world datasets and important results have been obtained.

Agriculture field provides many potential areas where many machine learning techniques are necessary for accurate and ultimate predictions. Prediction of rice produce is a challenging task for all farmers and there does not survive any standard framework for it. Uraiwan Inyaem [6] has proposed one machine learning technique by comparing decision tree and neural network techniques for increasing the rice production and income of farmers.

Cross validation is primarily used for controlling the generalization errors in the case of small datasets. In many contexts and real world situations the proposed machine learning algorithm must be compared with many competing and state-of-the-algorithms by using many standard datasets. Usage of many algorithms, folds, datasets, parameters, repetitions, comparisons and many other settings is a complex task. Author proposed a new protocol for drastic reduction of many potential operations of machine learning techniques. Bengio and Grandvalet [7] have studied generalization performance details of the k-fold cross validation technique by analyzing the Eigen values decomposition of the covariance matrix of errors.

Artificial intelligence applications are increasing day by day and Meta learning is one of the advanced fields of artificial intelligence. Many automatic learning algorithms are used for acquiring learning experience. Cross validation is one application of Meta learning for selecting the best learning algorithm. Cross validation is a time consuming process particularly in financially related applications involving large datasets. Xiang et al. [8] proposed two asymptotic Meta learning algorithms for cross validation and also experimentally proved that the proposed method has resulted significant improvements.

Test data accuracy of a machine learning model is not sufficient in determining the performance of the model and sometimes mere test accuracy parameter is insufficient for finding over fitting occurrence of the model. Cross validation is a technique that will be used for finding effective evaluation details of the models. In this paper Valier [9] experimentally said that cross validation is a potential technique for performance evaluation of learning models.

In machine learning the real or actual or original performance of a prediction algorithm is generally calculated in terms of calculated error. In cases where there is no method for finding exact prediction error then better to find an appropriate and approximate real estimator of the true prediction error. C. Chen et al. [10] have analyzed different statistical properties of the k-fold cross validation technique and experiments are conducted for comparing two prediction error finding algorithms. Also sensitivity results corresponding to the changes in the training datasets are also clearly observed through experiments.

The popular technique, k-fold cross validation, is normally used for finding the success rate of classification models. Only a few members have tried to find the best k value in k-fold cross validation. How to select the best k value and how k value affects the validation outputs is still an ongoing hot research topic. Marcot and Hanea [11] experimentally tried to demonstrate how validation results were affected in discrete Bayesian networks by considering dataset sizes, model structures and data dependencies. All these experiments were conducted by taking the known statistical properties such as variance and collinearity with different k values.

A desired model selection and evaluation is generally carried out by cross validation technique and in particular it is used in decision tree classification. In the case of general k-fold cross validation typical recommended value of k is 10 but it is not true in all cases. Karkkainen [12] has conducted many experiments on Multilayer Perceptron neural networks for finding the characteristics of k-fold cross validation method.

Creating a model with good generalization performance is the fundamental characteristic property in supervised learning. The given dataset must be divided intelligently for finding the optimal generalized model. Authors, Yun and Goodacre [13] have conducted many

experiments by taking many datasets and by using different data splitting techniques for splitting each dataset into training and validation datasets. The experimental results have shown that the size of data decides the generalization performance of validation datasets.

Performance of the classifier model is computed based on the prediction error and this error is estimated approximately instead of computing directly in real environment. Bias and variance parameters are used for error estimation. Rodriguez et al. [14] carried out experiments on two classifier models using different parameters including sample sizes, number of folds, and training datasets.

Tsung [15] introduced independent assumptions for applying k-fold cross validation and different contexts are given and the explained how these assumptions are applicable and useful. The definitions for various assumptions are clearly explained with some supports.

For large samples and folds there exist quadratic relationship between cross validation variance and its accuracy. Cross validation is frequently used for finding the generalization capability of the learning model. Gaoxia and Wen [16] thoroughly studied the relationships between variance and other factors such as data size, number of folds and number of runs and reported that error estimation can be decreased by decreasing the variance.

Dehghni et al. [17] conducted experiments on various datasets and experimentally proved that k-fold cross validation method increases the model performance accuracy when overlapping sub datasets are used in experiments and they suggested that overlapping sub datasets are not worth full due to their extra computational effort.

III. PROBLEM DEFINITION

In many machine learning applications k-fold cross validation is commonly used for estimating the model prediction accuracy and in particular in finding the estimated prediction accuracy of the decision tree model. Earlier researchers have predominantly decided that good k value in k-fold cross validation might be 10 or 20. In the present paper it is argued, discussed and explained that optimal k value (k = 10 or k = 20) is not always correct in k-fold cross validation in all applications. Optimal k value is dependent on many factors such as parameters like training data accuracy, data size, and number of classes in the training dataset and so on. In general, in many real time applications, actual k value in k-fold cross validation should have not been taken randomly as 10 or 20. In fact optimal k value must be determined only experimentally because each training dataset has got its own distinct and correct k value that is useful for that dataset.

Here the actual problem is how to find the correct k value for the given training dataset without always blindly assuming or take it for granted that k =10 or k = 20 or even directly taking any random value of k.

IV. EXPERIMENTAL RESULTS

In order to find optimal k value in k-fold cross validation many live experiments are conducted by employing many of the real world and realistic UCI machine learning standard datasets. Based on very close observations of the experimental results some of the useful conclusions and contributions are stated. Also these experimental results can be taken as the bench mark k values in many of the real and actual applications of these training datasets. Important conclusions of experimental results are:

1. Each training dataset has its own particular and distinct or unique k-value that must be computed through experimentation only but not randomly taken.

2. In k-fold cross validation random k-values should not be taken into consideration in all situations.

3. Usually, though the preferred values of k are 10 or 20, that is k = 10 or k = 20 in many cases but it should not be taken in a normal manner of take it for granted fashion without proper experimentation.

4. In the future, experimentally determined k values in k-fold cross validation can be standardized as the bench mark values of the k in k-fold cross validation.

S. No	K in k- Folds	Classification Accuracy
1	2	94.6666
2	3	94.6666
3	4	95.9459
4	**5**	**97.3333**
5	6	97.3333
6	7	97.2789
7	8	97.2222

8	9	97.2222
9	10	97.3333
10	15	97.3333
11	20	97.2789

Table-1 IRIS Dataset folds

From the Table-1, IRIS Dataset folds, it is clear that the optimal number of k value in k-fold cross validation for Iris training dataset is 5 as highest classification accuracy is achieved at that point in the very first time. IRIS dataset is very simple and very popular dataset in the machine learning repository of standard datasets.

S. No	K in k- Folds	Classification Accuracy
1	2	77.1028
2	3	82.1596
3	4	85.8490
4	5	89.5238
5	**6**	**92.8571**
6	7	92.3809
7	8	91.8269
8	9	92.7536
9	10	92.3809
10	15	87.14285
11	20	91.9047

Table-2 GLASS Dataset folds

From the Table-2, GLASS Dataset folds, it is clear that the optimal number of k value in k-fold cross validation for Glass training dataset is 6 because highest classification accuracy is achieved at that point. In the table classification accuracy is decreasing in the next increasing order of k values.

S. No	K in k- Folds	Classification Accuracy
1	2	90.2857
2	3	93.4472
3	4	83.0459
4	5	94.5714
5	6	94.5402
6	7	95.1428
7	8	83.7209
8	**9**	**98.8603**
9	10	83.1428
10	11	83.5777
11	12	83.3333
12	13	83.1908
13	14	83.1428
14	15	83.7681
15	20	83.5294

Table-3 IONO training Dataset

From the Table-3, IONO Dataset folds, it is clear that the optimal number of k value in k-fold cross validation for Iono training dataset is 9. Here up to 9 folds classification accuracy increases steadily and when k = 10 folds the classification accuracy starts to decrease. So, only the maximum classification accuracy value at the very first time is considered and it is taken as the value of k in k-fold cross validation.

S. No	K in k- Folds	Classification Accuracy
1	2	94.3661
2	3	95.4144
3	4	95.9507
4	5	95.5752
5	6	96.6312
6	**7**	**98.2363**
7	8	97.7112
8	9	97.7072
9	10	97.3214
10	11	97.3214
11	15	98.0180

12	20	98.0357

Table-4 BREAST Training Dataset

From the Table-4, BREAST Dataset folds, it is clear that the optimal number of k value in k-fold cross validation for Breast training dataset is 7. Up to 7 folds classification accuracy increases slowly and when k = 8 folds the classification accuracy starts to decrease. After 7-folds there are ups and downs in the classification accuracy. Though larger values of k also producing high classification accuracy results those k values are not taken into consideration and as result k should be as small as possible and it should be realistic also.

S. No	K in k- Folds	Classification Accuracy
1	2	75.4137
2	3	78.9598
3	4	75.9478
4	5	74.3195
5	6	80.0236
6	7	78.09523
7	8	79.2857
8	9	79.0780
9	10	79.8809
10	**15**	**80.4761**
11	20	79.8809

Table-5 VEHICLE Training Dataset

From the Table-5, VEHICLE Dataset folds, it is clear that the optimal number of k value in k-fold cross validation for Vehicle training dataset is 15. Up to 15 folds classification accuracy increases slowly and when k = 8 folds the classification accuracy starts to decrease. After 15-folds, classification accuracy decreases.

S. No	K in k- Folds	Classification Accuracy
1	2	93.7662
2	3	95.0649
3	4	95.4506
4	5	95.9740
5	6	96.3203
6	7	96.3586
7	8	96.3897
8	9	96.4762
9	**10**	**96.9876**
10	15	94.1298
11	20	94.1298

Table-6 SEGMENT Training Dataset

From the Table-6, SEGMENT Dataset folds, it is clear that the optimal number of k value in k-fold cross validation for Segment training dataset is 15. Up to 15 folds classification accuracy increases slowly and when k = 8 folds the classification accuracy starts to decrease. After 15-folds, classification accuracy decreases.

S. No	K in k- Folds	Classification Accuracy
1	2	94.1176
2	3	93.9393
3	4	93.7500
4	5	94.2857
5	6	94.2857
6	7	94.2857
7	8	93.7500
8	9	93.9393
9	10	93.9393
10	11	93.9393
11	**12**	**97.0588**
12	13	97.0588
13	14	97.0588
14	15	97.0588
15	16	97.0588

Table-7 INCOME Training Dataset

From the Table-7, INCOME Dataset folds, it is clear that the optimal number of k value in k-fold cross validation for Income training dataset is 12. Up to 11 folds, classification accuracy exhibits ups and downs and optimal value is achieved at k = 12 folds, then afterwards classification accuracy remains constant in all the further folds.

S.No	Dataset Name	Dataset Size	K in k-fold	Classification Accuracy
1	Iris	150	5	**97.3333**
2	Glass	214	6	**92.8571**
3	Iono	351	9	**98.8603**
4	Breast	569	7	**98.2363**
5	Vehicle	846	15	**80.4761**
6	Segment	2310	10	**96.9876**
7	Income	35	12	**97.0588**

Table-8 Training Datasets used in optimal k-fold determination

S.No	Dataset Name	K in k-fold
1	Iris	5
2	Glass	6
3	Iono	9
4	Breast	7
5	Vehicle	15
6	Segment	10
7	Income	12

Table-9 various datasets and their respective optimal k values in k-fold cross validation

For seven UCI machine learning datasets optimal and the most appropriate k values are experimentally computed and then tabulated in Table-9. In the future these values can be standardized in the process of determining the correct k value in the k-fold cross validation. Here k value is experimentally determined and k is dependent on many parameters such as number of classes in the training dataset, number of tuples, number of attributes, classification accuracy, technique that is used for k value determination and so on.

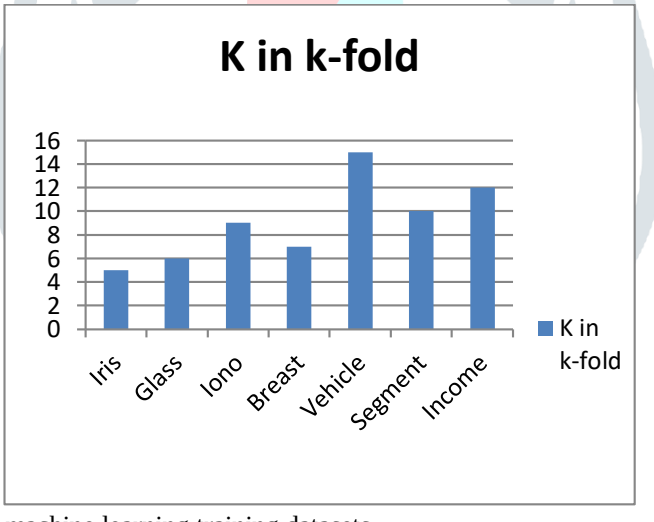

Fig-1 Optimal k values for various UCI machine learning training datasets

CONCLUSIONS

The present study experimentally explains applicability of new values of k in k-fold cross validation technique. The values k = 10 or k = 20 are the standard values which are used in applications. In the current study a new statement is made for selecting the k value. The selection of optimal k value is dependent on many factors and parameters of the training dataset. Also k value selection differs from one training dataset to another training dataset. In this paper actual k value selection is experimentally determined by computing classification accuracies of decision tree classification models. The various parameters that are to be considered for optimal k value selection are dataset size, number of distinct class labels, number of selected folds, and the size of the final test dataset and so on. In the future many more attractive features will be considered for selecting or determining the optimal k values in k-fold cross validation and in particular when the dataset sizes are small or very small or moderately small. In the future special methods and techniques will be considered for determining maximum k value in k-fold cross validation other than decision tree classification such as neural networks, support vector machines, k nearest neighbor classification, and so on.

REFERENCES

[1] Kaushika Pal and Biraj. V. Patel, "Data Classification with k-fold Cross Validation and Holdout Accuracy Estimation Methods with 5 Different Machine Learning Techniques", Published in: 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), Date of Conference: 11-13 March 2020, Date Added to IEEE Xplore: 23 April 2020

[2]Thineswaran Gunasegaran; Yu-N Cheah, "Evolutionary cross validation", Published in: 2017 8th International Conference on Information Technology (ICIT), Date of Conference: 17-18 May 2017, Date Added to IEEE Xplore: 23 October 2017

[3] P. Tamilarasi; R.Uma Rani, "Diagnosis of Crime Rate against Women using k-fold Cross Validation through Machine Learning", Published in: 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), Date of Conference: 11-13 March 2020, Date Added to IEEE Xplore: 23 April 2020

[4] Nurhayati; Indratmo Soekarno; Iwan K. Hadihardaja; M. Cahyono, "A study of hold-out and k-fold cross validation for accuracy of groundwater modeling in tidal lowland reclamation using extreme learning machine", Published in: 2014 2nd International Conference on Technology, Informatics, Management, Engineering & Environment, Date of Conference: 19-21 Aug. 2014, Date Added to IEEE Xplore: 19 January 2015

[5]Ramasubbu Venkatesh; Charles Rowland; Hongjin Huang; Olivia T. Abar; John Sninsky, "Robust Model Selection Using Cross Validation: A Simple Iterative Technique for Developing Robust Gene Signatures in Biomedical Genomics Applications", Published in: 2006 5th International Conference on Machine Learning and Applications (ICMLA'06), Date of Conference**:** 14-16 Dec. 2006, Date Added to IEEE Xplore: 26 December 2006

[6]Uraiwan Inyaem, "Construction Model Using Machine Learning Techniques for the Prediction of Rice Produce for Farmers", Published in: 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC), Date of Conference**:** 27-29 June 2018, Date Added to IEEE Xplore: 18 October 2018.

[7] Yoshua Bengio, Yves Grandvalet, "No Unbiased Estimator of the Variance of K-Fold Cross-Validation", Publication:The Journal of Machine Learning ResearchDecember 2004

[8] Haitao Xiang; Jianwu Lin; Chun-hung Chen; Ying Kong, "Asymptotic Meta Learning for Cross Validation of Models for Financial Data**",** *IEEE Intelligent Systems* (IF 3.210) Pub Date : 2020-02-11,*DOI:* 10.1109/mis.2020.2973255

[9] Valier A. (2020) "The Cross Validation in Automated Valuation Models: A Proposal for Use". In: Gervasi O. et al. (eds) Computational Science and Its Applications – ICCSA 2020. ICCSA 2020. Lecture Notes in Computer Science, vol 12253. Springer, Cham. https://doi.org/10.1007/978-3-030-58814-4_45

[10] Cuixian Chen, Yishj wang, Yaw Chang, and Karl Ricanek, "Sensitivity Analysis with Cross-Validation for Feature Selection and Manifold Learning", International Symposium on Neural Networks, ISNN 2012: Advances in Neural Networks – ISNN 2012 pp 458-467

[11] Marcot B.G., Hanea, A.M. "What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis?.", Computational Statistics (2020). https://doi.org/10.1007/s00180-020-00999-9., Published: 13 June 2020.

[12] Tommi karkkainen, "On Cross Validation for MLP Model Evaluation", Joint IAPR international workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)", 2014 pages 291-300.

[13]Yun Xu, R Goodacre, "Sampling for Estimating the Generalization Performance of Supervised Learning", Journal of Analysis and Testing, pp, 249 – 262 (2018),

[14] Juan D. Rodriguez, Aritz Perez, Jose A. Lozano, "Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation", Publication: IEEE Transactions on Pattern Analysis and Machine IntelligenceMarch 2010 https://doi.org/10.1109/TPAMI.2009.187

[15] Research article, Tzu -Tsung Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation", Publication: Pattern RecognitionSeptember 2015.
https://doi.org/10.1016/j.patcog.2015.03.009

[16] Gaoxia Jiang, Wenjian Wang, "Error estimation based on variance analysis of k-fold cross-validation", Publication: Pattern RecognitionSeptember 2017 https://doi.org/10.1016/j.patcog.2017.03.025

[17] Akbar Dehghni, Tristan Glatard, Emad Shihab, "Subject cross validation in Human activity Recognition", Cornell University, submitted on 4th April 2019, lastrevised 9th April 2019. arXiv:1904.02666