Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2176 – 2181
Received 05 March 2021; Accepted 01 April 2021.

# The Hybrid Feature Selection Method to Recognize Driver's Inattention Issue

**J. Mary DallfinBruxella[1] , Dr. J.K.Kanimozhi[2]**

*1 Ph.D(PT) Research Scholar, Department of Computer Science*
*Periyar University*
*Salem, India*

*2 Assistant Professor, PG & Research Department of Computer Science*
*Sengunthar Arts and Science College*
*Tiruchengode, India*

**Abstract**

Driver inattention detection has gained more attention in recent decades and the literature shows evidences of various Driving Monitoring and Assistance Systems. Those systems are aimed to assist the drivers to improve their performance and to avoid road accidents. The driving monitoring systems keep on the driving position of a driver and to deliver required support for comfortable and safe driving. A classification technique along with more number of features would be a comprehensive standard in this backdrop. Nevertheless, if the count of features are extremely high subsequently the intricacy of training phase would upsurge over and above there is a venture that the classification method might be contrary. This research work proposes a hybrid metric to estimate more robust feature score for preliminary feature selection. Moreover, the metric could be used to reduce the feature space both in horizontal as well as in vertical direction. The same metric has been used to identify the relevant features as well as to remove the redundant samples. The performance of the feature selection step is evaluated with Support Vector Machine (SVM) classifier.

***Keywords****Driver inattention, feature selection, PCC, hybrid metric, SNR*

## INTRODUCTION

It is essential to recognize the pertinent features, named as feature selection or else dimension reduction techniques.In recent years, various feature selection algorithms have been proposed. All of these feature selection methods can be disintegrated into three categories: filter, wrapper and hybrid methods. The hybrid methods effort to take benefit of the filter and wrapper methods by employing their complementary potencies. Hybrid techniques are usually a combination of filter and wrapper techniques and are designed to trade the accuracy with the computational speed by applying a wrapper technique to only those subsets that are preselected by the filter technique. The strategies used for searching the feature space in hybrid techniques are very different.

Choosing an appropriate feature selection strategy depends on the application. Though the feature selection algorithms are efficient, a preliminary feature selection step could reduce their computation cost further. The preliminary selection step is mostly based on filter approach, to make it simple. A feature score or ranking metric is estimated to measure the feature relevance, and the preliminary selection is made by choosing top 'n'

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2176 – 2181
Received 05 March 2021; Accepted 01 April 2021.

features from the complete feature set. Numerous feature score metric has been proposed in the literature as some of them are discussed in the following section. However, they are very much application dependent. This research work proposes a hybrid metric to estimate more robust feature score for preliminary feature selection. Moreover, the metric could be used to reduce the feature space both in horizontal as well as in vertical direction. The same metric has been used to identify the relevant features as well as to remove the redundant samples. The performance of the feature selection step is evaluated with Support Vector Machine (SVM) classifier.

## REVIEW OF LITERATURE

Kudo&Sklansky (2000) presented a comparative study between large scale feature selection algorithms. A degradation parameter based preliminary feature selection is implemented first to reduce the feature set, then the SFS algorithm is applied for further feature selection. The investigation results demonstrated the choice of feature selection algorithm based on the actual feature dimension. Goh et al., (2004) offered a preliminary feature selection method with Pearson Correlation Coefficient (PCC) and Signal-to-Noise Ratio (SNR) metrics, and reported greater classification accuracy with Gene expression data. Li et al., (2009) evaluated the effect of preliminary feature selection with microarray datasets. The t-statistic based ranking is used as a preliminary feature selection step to select the top 'l' genes and experimental reports shown that the preliminary step has strong influence on improving the classification accuracy significantly.

Xie et al., (2016) proposed a non-parametric statistical measure, Wilcoxon Signed-rank test is performed to rank the features first, then the top k-features are preliminarily chosen for feature selection. The experimental results indicate the significance of feature ranking. Ramos-González et al., (2013) offered a novel Case based Reasoning framework with gradient boosting based feature selection and applied to the task of lung cancer subtype classification. Here a non-parametric Mann-Whitney test is used to perform preliminary feature selection.

## HYBRIDMETRICS FOR FEATURE SELECTION

A combination of 3 metrics are used to filter the relevant features, these three metrics are chosen based on the literature study. The metrics are given as

(i) Pearson Correlation Coefficient (PCC) – Linear correlation coefficient is a measurement of the strength of a linear relationship between a dependent variable (i.e. the output class, $y$) and the independent variable (i.e. the sample, $x$).

$$PCC = \frac{\sum(x-x_{mean})(y-y_{mean})}{\sigma_x \sigma_y} \quad (3.1)$$

When $x$ increases and if $y$ also tends to increase or decrease, there is a mathematical linear dependency between $y$and $x$. The calculated $PCC$ gives a quantitative idea of the dependency. The correlation value ranges from –1 to 1. A value of 0 suggests no linear correlation, while values nearer to –1 or 1 means negatively or positively

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2176 – 2181
Received 05 March 2021; Accepted 01 April 2021.

correlated. PCC is for bivariate analysis, and provides a quick way to estimate linear relationship for data that has a normal distribution. However, for large data set, the computation time to calculate the PCC matrix is very long. Goh et al., (2004) proposed a novel way to calculate the matrix in a shorter time.

(ii) Signal-to Noise Ratio (SNR) is a calculated ranking number for each variable to define how well this variable discriminates two classes. The following formula is used:

$$SNR = (\mu_{class\,1} - \mu_{class\,2})/(\sigma_{class\,1} - \sigma_{class\,2}) \quad (3.2)$$

(iii) The Information Gain Ratio (IGR) is estimated as

$$IGR(D, C) = \frac{IG(D,C)}{H(D)} \quad (3.3)$$

For a dataset with the dimension $m \times n$, these six metrics are measured for each attribute, then the top $n'$ number of features are identified for each metric individually, where $n' = 0.75 * n$, the top 75% of the features having better relevance score are chosen from the original feature space. Union of these individual feature set is estimated to construct the first level feature subset. This step reduces the feature space respective to column dimension, known as dimensionality reduction in horizontal direction. In the next step, these metrics are estimated for each sample in the dataset, and the samples having same values are considered as redundant and the dataset dimension is reduced in vertical direction too. The following pseudocode illustrate the proposed preliminary feature selection algorithm.

## RESULTS AND DISCUSSION

The execution of the proposed preliminary feature selection method is examined with the driving dataset called Ford's stay alert.In Ford's stay alert driver's dataset, every trail is signifying a consecutive data, documented at each 100ms throughout the driving period on the highway. The sample comprises of hundred members of diverse age group, sexual category & racial credentials. The dataset is described as follows:

- The initial column indicates the ID for the trial
- The second column represents a sequential number within one trial ID
- The third column is the decision attribute to represent drivers inattention (0 – distracted, 1 – alert)
- The subsequent 8 columns exemplifies physiological data
- The next 11 columns represent environmental data
- The next 11 columns represent vehicular data

Around 33 columns of data gathered from 610 drivers & 1210 examinations individually. Altogether the dataset consumes facet of 738100 instances through 33 events, everywhere the initial 2 columns might be discounted as they sustain the serial numbers. Thus, the dataset facet is abridged to 738100×31. Further, the

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2176 – 2181
Received 05 March 2021; Accepted 01 April 2021.

dataset is split into 2 groups (i.e.,) five hundred and ten drivers' examinations for training & the remaining hundred drivers' examinations for testing.

Table 3.2 presents the number of selected attributes from all the ford's stay alert dataset after filter based preliminary feature selection. Comparatively the hybrid metric selects similar features as illustrated in Table 3.3.

Table 3.2 Number of preliminary selected attributes from Driver's Dataset

| **Feature Selection Methods** | **Hybrid Metric** | **PCC** | **SNR** | **IGR** |
|---|---|---|---|---|
| **Ford's Stay Alert** | 24 | 22 | 18 | 28 |

| **Feature Selection Methods** | **Ford's Stay Alert (738100 × 31)** |
|---|---|
| **PCC** | 712184 |
| **SNR** | 716395 |
| **IGR** | 708201 |
| **Hybrid Metric** | 689380 |

Table 3.3 List of preliminary selected attributes from Ford Stay's Dataset

| **Dataset** | **List of Selected Attributes** |
|---|---|
| **Ford's Stay Alert (25)** | 1, 2, 3, 4, 5, 7, 9, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 25, 27, 28, 29, 30, 31 |

Table 3.4 vertically reduced dimension

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2176 – 2181
Received 05 March 2021; Accepted 01 April 2021.

For the vertical dimensionality reduction, the reduced dimension is reported in Table 3.4, demonstrate that the hybrid relevance score is achieving more meaningful reduction than the other metrics. Though the selected number of features from the hybrid metric is slightly higher than the results from some other feature relevance metric, the classification results shown in Table 3.5 indicates that the hybrid feature score metric improves the accuracy.

Table 3.5 Classification performance on preliminary selected attributes

| **Feature Selection Methods** | **Ford's Stay Alert** |
|---|---|
| **PCC** | 0.8998 |
| **SNR** | 0.8781 |
| **IGR** | 0.8821 |
| **Hybrid Metric** | 0.9182 |

## CONCLUSION

A novel hybrid feature relevance score metric is proposed for preliminary feature selection in driver inattention detection application. The hybrid score is based on Pearson Correlation Coefficient (PCC), Signal-to-Noise Ratio (SNR), Information Gain Ratio (IGR) metrics. The top 75% of features are chosen after each metric, then the subsets are combined together to derive the preliminarily reduced feature set. Further, the redundant samples from the dataset are identified by the applying each relevance metric at row level, and the multiple samples with the same metric are removed to avoid the duplicate records. The classification performance of the proposed feature selection is evaluated with a Support Vector Machine (SVM) classifier, and demonstrated that the preliminary feature selection achieves better classification significantly.

## REFERENCES

1. Åkerstedt, T., &Gillberg, M. (1990). Subjective and objective sleepiness in the active individual. International Journal of Neuroscience, 52(1-2), 29-37.
2. deNaurois, C. J., Bourdin, C., Stratulat, A., Diaz, E., &Vercher, J. L. (2017). Detection and prediction of driver drowsiness using artificial neural network models. Accident Analysis & Prevention.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2176 – 2181
Received 05 March 2021; Accepted 01 April 2021.

3. Du, Y., Wang, Y., Huang, X., & Hu, Q. (2018). Driver State Analysis Based on Imperfect Multi-view Evidence Support. Neural Processing Letters, 48(1), 195-217.
4. Gjoreski, M., Gams, M. Ž.,Luštrek, M., Genc, P., Garbas, J. U., & Hassan, T. (2020). Machine Learning and End-to-End Deep Learning for Monitoring Driver Distractions From Physiological and Visual Signals. IEEE Access, 8, 70590-70603.
5. Goh, L., Song, Q., &Kasabov, N. (2004, January). A novel feature selection method to improve classification of gene expression data. In Proceedings of the second conference on Asia-Pacific bioinformatics-Volume 29 (pp. 161-166). Australian Computer Society, Inc..
6. González-Ortega, D., Díaz-Pernas, F. J., Antón-Rodríguez, M., Martínez-Zarzuela, M., &Díez-Higuera, J. F. (2013). Real-time vision-based eye state detection for driver alertness monitoring. Pattern Analysis and Applications, 16(3), 285-306.
7. Hari, C. V., &Sankaran, P. (2017, August). Embedding vehicle driver face poses on manifolds. In 2017 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES) (pp. 1-5). IEEE.
8. Kudo, M., &Sklansky, J. (2000). Comparison of algorithms that select features for pattern classifiers. Pattern recognition, 33(1), 25-41.
9. Li, G. Z., & Zeng, X. Q. (2009). Feature selection for partial least square based dimension reduction. In Foundations of Computational Intelligence Volume 5 (pp. 3-37). Springer, Berlin, Heidelberg.
10. Xie, J., Wang, M., Zhou, Y., & Li, J. (2016, August). Coordinating discernibility and independence scores of variables in a 2D space for efficient and accurate feature selection. In International Conference on Intelligent Computing (pp. 116-127). Springer, Cham.