01 Jul 2020

Improved Classification Of Medical Data Using Meta-Best Feature Selection

Matthew Chaplin

Jacob Grubb

Thomas Clifford

Justin Bruce

et. al. For a complete list of authors, see https://scholarsmine.mst.edu/ele_comeng_facwork/4816

Recommended Citation

M. Chaplin et al., "Improved Classification Of Medical Data Using Meta-Best Feature Selection," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pp. 5602 - 5605, article no. 9175289, Institute of Electrical and Electronics Engineers, Jul 2020.
The definitive version is available at https://doi.org/10.1109/EMBC44109.2020.9175289

Improved Classification of Medical Data Using Meta-Best Feature Selection

Matthew Chaplin*, Jacob Grubb*, Thomas Clifford, Justin Bruce,
Tayo Obafemi-Ajayi, and John Matta

***Abstract*— Feature selection provides a useful method for reducing the size of large data sets while maintaining integrity, thereby improving the accuracy of neural networks and other classifiers. However, running multiple feature selection models and their accompanying classifiers can make interpreting results difficult. To this end, we present a data-driven methodology called Meta-Best that not only returns a single feature set related to a classification target, but also returns an optimal size and ranks the features by importance within the set. This proposed methodology is tested on six distinct targets from the well-known REGARDS dataset: *Deceased*, *Self-Reported Diabetes*, *Light Alcohol Abuse Risk*, *Regular NSAID Use*, *Current Smoker*, and *Self-Reported Stroke*. This methodology is shown to improve the classification rate of neural networks by 0.056 using the ROC Area Under Curve metric compared to a control test with no feature selection.**

I. Introduction

Collection of large amounts of electronic medical and biological data has made the use of big data analysis techniques critical for smart healthcare. A prominent issue when analyzing large datasets is an over-abundance of features. Feature selection is a process of selecting a representative subset of data, which, when used with machine learning techniques, aims both to increase prediction accuracy and to decrease model training time. Studies have found feature selection to be worth the effort [1] with regards to identifying trends in big data and forming meaningful conclusions on the attributes within. In this paper we apply a data-driven methodology using feature selection and neural networks. We improve upon previous feature selection methods and demonstrate results on a large medical dataset.

The REasons for Geographic And Racial Differences in Stroke (REGARDS) [2] dataset is the result of a well-known study involving stroke prevalence. It consists of medical, social, and economic data on 30,239 individuals collected between the years of 2003 and 2007. Here we provide an analysis of over 400 cases of feature selection and classification for biomarkers in the REGARDS dataset and examine the improvement in classification via neural networks after feature set reduction. We also introduce a novel ensemble method called Meta-Best intended to discern optimal feature subsets for a given classification target. Use of the data involving human subjects described in this paper was approved by the SIUE Institutional Review Board.

*Matthew Chaplin and Jacob Grubb contributed equally to this work.

Matthew Chaplin, Jacob Grubb, Thomas Clifford, Justin Bruce and John Matta are with the Computer Science Department, Southern Illinois University Edwardsville, Edwardsville, IL 62026 USA (e-mail:jmatta@siue.edu).

Tayo Obafemi-Ajayi is with the Engineering Program, Missouri State University, Springfield, MO 65897 USA (e-mail: TayoObafemiAjayi@MissouriState.edu).

II. Methods

A. Data Acquisition and Preparation

The cleaning and normalization of the REGARDS dataset was completed as described in detail in [3]. To test the performance of feature selection on a derived variable, a new feature was created: *Light Alcohol Risk*. This variable represents the subject's risk of alcohol abuse, as defined by the National Institute on Alcohol Abuse and Alcoholism (NIAAA) [4]. The new feature was created based the NIAAA definition, the biological sex of the subject, and the number of alcoholic drinks consumed per week. The feature *Alcoholic Drinks / Wk* was removed.

B. Feature Selection Methods

We employed two different families of feature selection: Sequential Forward Selection [5] and Correlation-based Feature Selection Subset Evaluation (CFS Subset Evaluation) [6]. Sequential Forward Selection is a greedy algorithm that works by taking a single feature from the set, adding it to the subset of selected features, and testing the subset using a classifier (or wrapper). By repeating this process the algorithm is able to build up the best subset of selected features for that specific classifier. Within Sequential Forward Selection, we test three wrapper methods: Logistic Regression [7], Naïve Bayes [8], and Random Forest [9].

CFS Subset Evaluation, by contrast, seeks to find subsets of attributes that are highly correlated with the target feature, but include features that have low correlation amongst themselves. Thus CFS Subset Evaluation is a grading criteria for different search algorithms. We test Best First [10], Evolutionary Search [11], and Minimum Redundancy Maximum Relevance (mRMR) [12] for our CFS Subset Evaluation.

The Sequential Forward Selection was performed using the Scikit-Learn Python package [13], while the CFS Subset Evaluation utilized Weka [14].

An important distinction between these two families of feature selection is that while CFS Subset Evaluation automatically chooses the size of the feature subset, Sequential Forward Selection requires the size of the feature subset, k, to be chosen before execution. We test feature sets of size $k = 5,\ 10,\ 15,\ 20,$ and 65 for each targeted feature. This results in 5 sets of selected features per classification target using Sequential Forward Selection, as compared to one set of features per target when using CFS Subset Evaluation.

978-1-7281-1990-8/20/$31.00 ©2020 IEEE

C. Performance Testing: Neural Networks

For each target feature, a feed-forward neural network was first run without feature selection in order to establish a control case. Then a neural network with the same hyperparameters was run using the feature subset produced by the feature selection method.

To rate the performance of the feature selection methods, we use the Receiver Operating Characteristic Area Under Curve (AUC) produced by these neural networks. All neural networks were run with the following hyperparameters: a learning rate of 0.3, a momentum of 0.2, and a training time of 500 epochs. Each network had one hidden, fully-connected layer, and no dropout was used. To ensure the robustness of our neural networks, 10-fold cross validation was employed.

D. Meta-Best Feature Selection

The *Meta-Best* feature selection method is used to create one optimal feature subset per classification target using the best input from each of the previous feature selection methods. This method does not require a feature subset size as an initial parameter, and allows for the combining of inputs from feature selection methods that may or may not require such a parameter. The method builds upon the work of [15] by adding a weighting system which takes into account the performance of each feature selection method, rather than simply scoring the goodness of a feature via how often it appears in a feature subset. Each feature selection method is assigned a score S from 0 to 1 based on its AUC gain using the following formula:

$$S = b_{method}/b_{target} \quad (1)$$

where b_{method} is the best AUC gain for the target feature generated by the given feature selection method, and b_{target} is the best AUC gain generated for the target feature *across all methods*. The *goodness* G of a feature is then graded using the formula:

$$G = (m)/((S_1 * a_1) + (S_2 * a_2) + ... + (S_N * a_N)) \quad (2)$$

where m is the number of times a feature appears across all subsets, a_i is a binary variable indicating whether or not the feature appeared in the method's specific subset, and N is the number of feature selection methods drawn from (in our case, 6 methods). This dynamic approach allows Meta-Best feature selection to choose the size of its feature set, much like CFS Subset Evaluation.

The goodness score, G, which has been generated by Equation 2 is then compared with a cutoff threshold, t, which is set manually, to indicate whether or not the feature is good enough to be included in the subset produced by Meta-Best; i.e. if a feature's G value was below the t value for that test, it was removed from the feature subset. For our experiments, the cutoff threshold range was 0 to 4, with a step of 0.1 between each case, giving us a maximum of 41 test cases per target feature (note that some target features had no features selected for very high thresholds). These numbers were chosen because features with a G score below 0 would indicate a negative effect on the overall AUC gain, and 3.924 was the highest score of any selected feature, as discussed in Section III-C. Once all possible test cases for a target feature had been run, the feature subset with the highest AUC gain was considered to be the optimal feature subset, with greater t values being used to determine the winner in the case of a tie.

TABLE I: The AUC gain (indicated by AUC) and the local optimal number of features (indicated by k) produced by each algorithm for each target feature, with the target feature's globally optimal feature subset shown in bold.

	Deceased		SR Diabetes		Alc. Risk	
Algorithm	k	AUC	k	AUC	k	AUC
Best First	17	0.038	7	0.003	5	0.051
Evolutionary Search	31	0.013	30	-0.007	27	-0.012
Logistic Regression	***10***	***0.061***	***15***	***0.044***	5	0.062
mRMR	17	0.038	7	0.003	5	0.051
Naïve Bayes	10	0.059	15	0.036	***10***	***0.065***
Random Forest	20	0.048	20	0.028	10	0.038

	NSAID User		Curr. Smoker		SR Stroke	
Algorithm	k	AUC	k	AUC	k	AUC
Best First	7	0.054	12	0.033	26	0.014
Evolutionary Search	21	0.042	29	0.019	27	-0.02
Logistic Regression	***10***	***0.072***	***15***	***0.054***	10	0.053
mRMR	4	0.046	10	0.032	20	0.019
Naïve Bayes	10	0.065	15	0.042	***15***	***0.054***
Random Forest	15	0.015	65	-0.008	15	0.037

III. Results

A. Size of Feature Subsets

We tested the optimal feature subset size by comparing the AUC gain for each target's feature subsets against the size of these feature subsets k. Note that although multiple feature set sizes were tested, for brevity only the best AUC gain results are shown in Table I. It can be seen that strong AUC increases were produced by both Logistic Regression and Naïve Bayes.

B. Feature Selection Method Comparisons

To test the performance of feature selection methods, we compare the average AUC gain from neural network classifiers across all target features, displayed in Figure 1. For the case of Sequential Forward Selection, we include the performance of each subset for $k = 5$, 10, 15, 20, and 65 features chosen. For the Meta-Best feature selection method, all neural network AUC results generated for cutoff thresholds between 0 and 4 were averaged.

Of the non-ensemble feature selection methods, Logistic Regression performed best with an average AUC gain of 0.041, with Evolutionary Search performing worst with an average AUC gain of 0.006. The Sequential Forward Selection methods tended to perform more strongly than their CFS Subset Evaluation counterparts, with the average gain across all SFS methods being 0.030 compared to CFS Subset Evaluation's 0.023. This is in spite of the fact that difficult cases (such as feature sets containing 5 or 65 features) were included in the unweighted average.

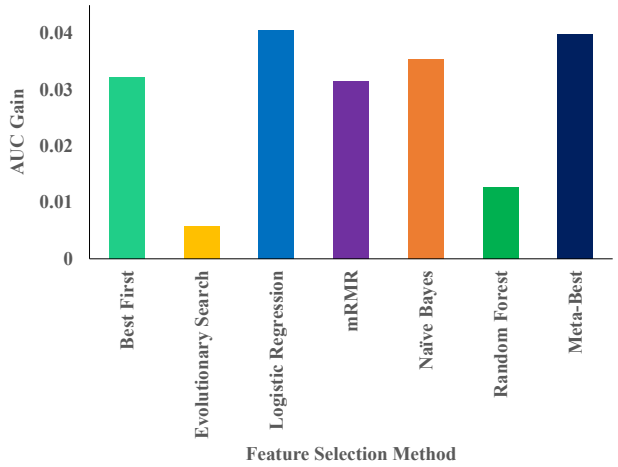

Fig. 1: The classification performance improvement of feature selection methods across all target features, as shown by the neural networks' average AUC gain over the control case. Note that in spite of decreasing classifier performance in fifty percent of cases, Evolutionary Search still provides a net increase across all target features. In addition to having nearly equal performance, Best First and mRMR produced similar sized feature sets with Best First choosing 1.8 more selected features on average.

Meta-Best performed slightly worse than Logistic Regression when averaged across all of its test cases, with an average AUC gain of 0.040. However, when its cutoff threshold is constrained to its optimal range (discussed in Section III-C, shown in Figure 2), Meta-Best improved to an AUC gain of 0.047.

C. Meta-Best Subset Results

As shown in Table II, the Meta-Best algorithm is appealing not only because of its high performance and ability to choose its own feature set size, but also because it returns a single optimal (or near-optimal) subset for a given target feature. The features of these subsets are ranked by their goodness score G from Equation 2, providing an advantage over Sequential Forward Selection.

As presented in Figure 2, comparing the Meta-Best AUC gain against its cutoff threshold t results in a rough negative parabolic curve. Optimal feature subsets were produced between threshold values of $t = 1.0$ and $t = 2.5$, inclusive. AUC gain (and AUC score) were noted to decay at any cutoff threshold higher than 2.5, most likely due to features introducing noise in the subset. The optimal feature subsets are marked in the figure by diamonds. In cases of ties, the case with the higher cutoff threshold was chosen, as this would produce the same classification rate with a smaller feature subset size.

For Meta-Best's optimal cases, it always produced a feature subset that was the same size or smaller than the best feature selection method, choosing an average of 2.33 fewer features for that target's feature subset. This was accomplished while performing within a range of +/- 0.005 AUC score of the best non-ensemble feature selection method.

TABLE II: Meta-Best's Chosen Feature Subsets For Each Target Feature

Current Smoker

Feature	*G*
Age	3.186
Body Mass Index	3.186
Self-reported Health	3.186
Education: College Plus	2.830
Heart rate	2.830
CESD	2.408
Income: Less than 20k	2.408
Light Alcohol Abuse Risk	2.408
Current Alcohol Use	1.982
Race	1.982
Heavy Alcohol Risk	1.834
HDL Cholesterol	1.630
Deceased	1.630
Has Health Insurance	1.408

Self-Reported Diabetes

Feature	*G*
Total Cholesterol	2.454
Insulin	2.431
Insulin Use	2.431
Self-reported Health	2.295
Glucose	2.295
Anti-hypertensive meds	2.295
Eleveated lipids	2.295
Current Alcohol Use	1.795
Waist circumference	1.660
Cystatin C	1.480
Triglycerides	1.480
Education: Some College	1.454
Biological Sex	1.454

Regular NSAID User

Feature	*G*
PCS-12: SF-12 Physical	3.875
Biological Sex	3.444
Race	2.861
Current Alcohol Use	1.903
Self-Reported Stroke	1.791
Atrial fibrilation	1.750
Body Mass Index	1.750

Light Alcohol Abuse Risk

Feature	*G*
Current Alcohol Use	3.924
HDL Cholesterol	3.924
Current Smoker	3.524
Biological Sex	2.540
Race	1.769

Deceased

Feature	*G*
Age	3.590
Cystatin C	3.590
Self-reported Health	3.590
Biological Sex	2.967
Heart rate	2.967
Current Smoker	2.754
PCS-12: SF-12 Physical	2.620
Albumin/Creatinine ratio	2.590

Self-Reported Stroke

Feature	*G*
Deceased	3.277
Fall in the Past Year	2.925
Self-reported Health	2.907
Anti-hypertensive meds	2.907
Regular Aspirin User	2.907
Reported TIA at Baseline	2.907
Current Alcohol Use	2.592
PCS-12: SF-12 Physical	2.222
Age	1.980
Regular NSAID User	1.980
Repair of aortic aneurysm	1.315
Income: Greater than 75k	1.296
Perceived Stress Scale	1.259

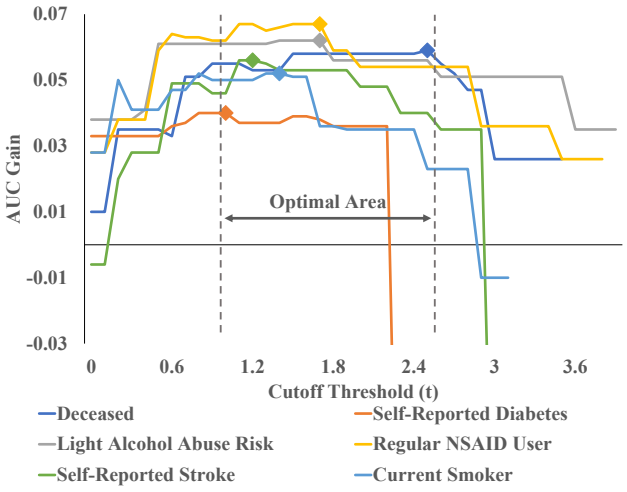

Fig. 2: The classification improvement versus cutoff thresholds of Meta-Best feature selection. The optimal cutoff threshold range is from 1.0 to 2.5, delineated with dashed lines. Optimal cutoff thresholds for a target feature are marked as diamonds. Some targets do not continue to the maximum value because no features scored high enough to form feature subsets for the given cutoff threshold t.

IV. Discussion and Conclusions

Results highlighted the tradeoffs between using Correlation-based Feature Subset Selection and Sequential Forward Selection; namely, that CFS Subset Selection allows for reasonable results with far fewer trials due to its automatic choice of feature subset size, while SFS provided better classification performance in exchange for the extra compute time to find its optimal feature subset size.

Of further interest were the actual feature subsets produced by Meta-Best. *Self-Reported Diabetes* tied for second longest feature list at 13 entries, in spite of the fact that it was the most easily classified feature – both in terms of control case AUC score and average AUC score after feature selection. We suspect that *Self-Reported Diabetes* is a feature largely unaffected by noisy values, due to having a larger number of potentially good features. This suspicion stems from the fact that it has the lowest maximum G value of any of our chosen target features at 2.454, suggesting that the component feature selection methods took different approaches on what features were used to classify it.

It is interesting to note which feature was chosen as the most relevant for each target class. *Deceased* was the most selected feature for *Self-Reported Stroke*, meaning that the model was "looking back in time" in a sense in order to attempt to classify the subject as a stroke victim. *Total Cholesterol* was considered more important than *Insulin* in determining whether or not a patient suffered from diabetes, while *Age* was considered the most important factor for both *Deceased* and *Current Smoker*. *Current Alcohol Use* was not automatically removed as a highly correlated feature for *Light Alcohol Abuse Risk*, meaning that enough people were responsible alcohol users within the scope of the survey that the algorithms were not able to classify those with abuse risk using this feature alone.

Out of 71 potential features, only 37 features were chosen across all of the feature subsets produced by Meta-Best feature selection, with six features showing up in at least half of the target feature subsets. *PCS-12: SF-12 Physical*, *Age*, and *Race* appeared in half of the feature subsets. *Biological Sex* and *Self-reported Health* appeared in four of the six lists, and *Current Alcohol Use* appeared in all of the feature subsets except *Deceased*, making it the most common feature selected across all targeted features.

A potential improvement to Meta-Best would be adding a search function to reduce the number of classifiers run. For instance, due to the roughly negative parabolic shape displayed by the AUC gains of the neural networks after Meta-Best Feature Selection in comparison to their cutoff thresholds, a hill-climbing algorithm could be implemented to approximate the optimal cutoff threshold. This would allow Meta-Best Feature Selection to perform much like Sequential Forward Selection, in that it would self-optimize for the best feature subset using a classifier, but would also have the ability to approximate the optimal feature subset size like Correlation-based Feature Subset Evaluation. Further work is needed to evaluate this approach.

In conclusion, we have shown that classification of target features is improved by feature selection, and have quantified this improvement across a range of feature selection methods and feature subset sizes. We have compared the methodology's ability to classify six distinct target features, and have provided an optimal feature subset for each of them using Meta-Best feature selection. We hope this methodology will be of use to those seeking to overcome the difficulties inherent in using big data for biomarker classification of healthcare data.

References

[1] A. Montillo and H. Ling, "Is feature selection worth the effort? assessing the impact on random forest and svm accuracy and computation time."

[2] V. J. Howard, M. Cushman, L. Pulley, C. R. Gomez, R. C. Go, R. J. Prineas, A. Graham, C. S. Moy, and G. Howard, "The reasons for geographic and racial differences in stroke study: objectives and design," *Neuroepidemiology*, vol. 25, no. 3, pp. 135–143, 2005.

[3] T. Clifford, J. Bruce, T. Obafemi-Ajayi, and J. Matta, "Comparative analysis of feature selection methods to identify biomarkers in a stroke-related dataset," in *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, 2019.

[4] "Drinking levels defined," Jan 2017. [Online]. Available: https://www.niaaa.nih.gov/alcohol-health/overview-alcohol-consumption/moderate-binge-drinking

[5] D. W. Aha and R. L. Bankert, "A comparative evaluation of sequential feature selection algorithms," in *Learning from Data*. Springer New York, 1996, pp. 199–206.

[6] M. A. Hall and L. A. Smith, "Feature subset selection: a correlation based filter approach," 1997.

[7] S. Menard, *Applied logistic regression analysis*. Sage, 2002, vol. 106.

[8] I. Rish *et al.*, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, 2001, pp. 41–46.

[9] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[10] L. Xu, P. Yan, and T. Chang, "Best first strategy for feature selection," in *[1988 Proceedings] 9th International Conference on Pattern Recognition*. IEEE, 1988, pp. 706–708.

[11] Y. Kim, W. N. Street, and F. Menczer, "Feature selection in unsupervised learning via evolutionary search," in *KDD*, 2000, pp. 365–369.

[12] M. Radovic, M. Ghalwash, N. Filipovic, and Z. Obradovic, "Minimum redundancy maximum relevance feature selection approach for temporal gene expression data," *BMC bioinformatics*, vol. 18, no. 1, p. 9, 2017.

[13] F. Pedregosa, Varoquaux, and et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[14] Z. Markov and I. Russell, "An introduction to the weka data mining system," in *ACM SIGCSE Bulletin*, vol. 38, no. 3. ACM, 2006, pp. 367–368.

[15] L.-R. Alejandro, M.-A. Marlet, M.-R. Gustavo Ulises, and T. Alberto, "Ensemble feature selection and meta-analysis of cancer mirna biomarkers," *bioRxiv*, 2018. [Online]. Available: https://www.biorxiv.org/content/early/2018/06/21/353201