Comparison of Various Feature Selection Techniques for Binary Classification Tasks

NDSU

UNIVERSIDAD DE CHILE

Andriy Tryshnivskyy, Simone Ludwig, Aaron Mackenzie

01. Introduction

In today's world data is used in almost every field. Data is collected by hospitals, governments, schools, and even farmers. Various processing methods are implemented in order to take the large quantities of information and turn them into results that can be comprehended. One such method is feature selection. For classification problems, feature selection can allow for the filtering of data in order to remove irrelevant information and achieve higher accuracy for classification models. For the project, various industry standard feature selection techniques have been implemented, including Information Gain, Forward Selection, Lasso Regression, and Chi-Squared with Simulated Annealing. In addition to this, a published research paper regarding the use of a variant of Particle Swarm Optimization for feature selection was analyzed and the strategy that was used was implemented and tested as well. All of these techniques were compared using statistical measures.

02. Related Works

Traditional PSO

In traditional PSO for feature selection, particles are initialized with the number of features being defined. Using the number of features, the position and velocity variables were created with random numbers between 0 and 1. The size of the position and velocity vectors was equal to the number of features defined at the start. Then for updating the position and score variables, you would only update them if the fitness value (accuracy) for a subset of features exceeded that of either the personal best subset of features or the global best subset of features.

PSO Variant

On the other hand, for the PSO for feature selection that was described in the Xue's paper, various initialization techniques and updating mechanisms were described. Within the paper, PSO(4-2) (PSO variant using mixed initialization as well as having the classification be the first priority when updating best scores and positions) was identified as the top performer when all of the variants were used, so this is the implementation that was used to perform the tests on the datasets that were given. For both traditional PSO and Xue's variants, KNN (n=5) was used as the classification model to run tests.

03. Methodology

Perform feature selection on each dataset and analyze the performance.

Performance Metrics

- **Accuracy**
- **Precision**
- **F1 Score**
- **Recall**
- **Friedman Test**
- **Nemenyi Test**

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

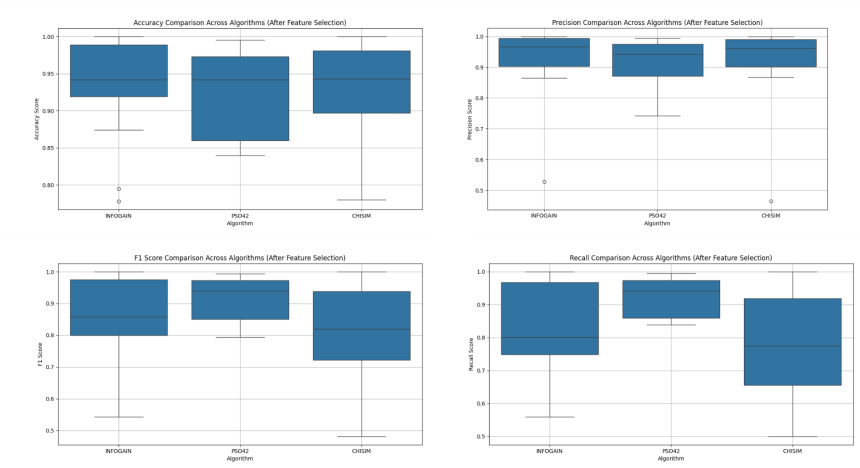
$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

04. Datasets

Data	Rows(Samples)	Columns(Features)	Class 1	Class 2
CHOL_EX	44	43,697	39	5
COAD_EX	522	37,677	509	13
HNSC_EX	564	35958	535	29
KICH_EX	91	43806	60	31
KIRC_EX	613	44909	609	4
KIRP_EX	322	44874	236	86
LIHC_EX	421	35924	322	99
LUSC_EX	553	44894	494	59
PRAD_EX	553	44824	472	81
STAD_EX	448	44878	358	80
THCA_EX	564	36120	380	184
UCEC_EX	588	36,086	345	243

05. Results



06. Analysis of Results

Friedman Test for Accuracy:

- Statistic: 4.87
- P-value: 0.0876

Friedman Test for Precision:

- Statistic: 7.48
- P-value: 0.0238
- Nemenyi Test Results for Precision:
  - Information Gain vs. PSO(4-2): p = 0.5597 (not significant)
  - Information Gain vs. Chi-Squared with Simulated Annealing: p = 0.2320 (not significant)
  - PSO(4-2) vs. Chi-Squared with Simulated Annealing: p = 0.0217 (significant difference)

Friedman Test for Recall:

- Statistic: 9.91
- P-value: 0.0070
- Nemenyi Test Results for Recall:
  - Information Gain vs. PSO(4-2): p = 0.1577 (not significant)
  - Information Gain vs. Chi-Squared with Simulated Annealing: p = 0.0062 (significant difference)
  - PSO(4-2) vs. Chi-Squared with Simulated Annealing: p = 0.4404 (not significant)

Friedman Test for F1 Score:

- Statistic: 5.39
- P-value: 0.0675

Upon analyzing the results, it can be concluded that there are no significant differences in accuracy and f1 scores for the feature selection methods. On the other hand, precision and recall do vary significantly. Based on these significant differences, the use cases for each feature selection technique can be defined. When high precision is needed, PSO(4-2) should be used over chi-squared with simulated annealing. When high recall is needed, information gain should be used as the feature selection method over chi-squared with simulated annealing.

07. Conclusion

The results from this study highlight the importance of choosing the appropriate feature selection technique based on the most relevant performance metric for a specific task. The analysis in this study involved evaluating traditional feature selection methods such as Information Gain and Chi-Squared with Simulated Annealing with a PSO variant in order to isolate the best feature selection technique in the context of highly dimensional data. The results state that accuracy and f1 did not differ significantly over the different methods, but when a certain performance metric needs to be prioritized, the specific algorithm that correlates to a greater difference in that metric should be used. In this case, Chi-squared with simulated annealing does not yield any significant advantages over the other feature selection methods. Overall, the study showcases the nuances that characterize feature selection and the careful analysis that needs to be done before choosing a method for the desired application. This work contributes to the ongoing efforts to increase performance and efficiency in the world of data analytics and machine learning, and the future work associated with this project would be to develop an original algorithm with increased performance compared to those that already exist.

08. References

B, H. N. (2020, June 1). Confusion matrix, accuracy, precision, recall, F1 score. Medium. https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd

Wallis, D. (2021, October 3). Comparing classifiers (Friedman and Nemenyi tests). Medium. https://medium.com/@diogeneswallis/comparing-classifiers-friedman-and-nemenyi-tests-32294103ee12

Xue, B. (2013, October 26). Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms. Applied Soft Computing. https://www.sciencedirect.com/science/article/abs/pii/S1568494613003128?via%3Dihub