Data Complexity and Meta-Learning - Health Insurance Dataset

Francisco da Ana (up202108762)
(Dated: November 10, 2024)

This report explores data complexity analysis and meta-learning concepts applied to a Health Insurance dataset. The analysis uncovers a careful observation of data-quality metrics and their implications in some classical classification algorithms.

I. INTRODUCTION

Data complexity and meta-learning are key concepts in modern machine learning. Meta-learning automates the learning process by understanding algorithm performance across tasks, while data complexity analysis quantifies dataset characteristics that affect algorithm performance. These complexity measures, such as class separability and feature distributions, serve as meta-features. In this report, we analyze a Health Insurance dataset using `problexity` to identify key complexity factors and their impact on classification paradigms.

II. HEALTH INSURANCE DATASET

This analysis continues the exploration of the Health Insurance dataset previously examined in our Data Profiling assignment. The dataset comprises about 72,000 instances with 15 features, where the binary target variable `health_ins` indicates whether an individual has health insurance or not. This dataset choice allows us to build upon our previous understanding while investigating its complexity characteristics for a binary classification problem.

III. DATA COMPLEXITY ANALYSIS

The dataset's complexity was analyzed using `problexity`, which provided measures across several categories: feature-based, linearity, dimensionality, class imbalance, neighborhood, and network characteristics. Figure 1 presents these measures graphically.

A. Key Findings

The analysis revealed several important characteristics of the dataset. High class separability (F1) suggests a clear distinction between insurance holders and non-holders, though moderate feature overlap (F2) introduces some ambiguity in class boundaries. The dataset's predictive potential is strong, with high single feature effectiveness (F3) and combined feature effectiveness (F4).

The low error distance (L1) and error rate (L2) indicate excellent linear separability, while the low non-linearity measure (L3) suggests stable decision boundaries.

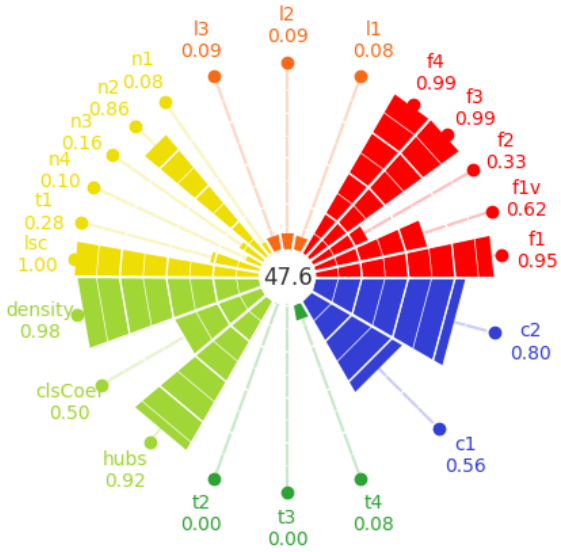

FIG. 1. Data Complexity Measures from `problexity`.

Dimensionality analysis shows significant redundancy, with a low dimensionality ratio (T4) and PCA revealing that one dimension captures 95% of the data variance. The very low T2 value (2×10^{-4}) suggests a favorable instance-to-feature ratio.

Class imbalance (C2) and moderate entropy (C1) are present, but local class separation (LSC) is strong, with minimal borderline points (N1). A high ratio of intra/extra class NN distance (N2) indicates that classes are quite close compared to the distances between different classes.

B. Implications for Classification

These complexity measures suggest that linear classifiers will perform well due to the strong separability and clear decision boundaries. Dimensionality reduction could improve model efficiency, and feature selection may help remove redundant variables. Class imbalance handling techniques may be necessary, though simple neighborhood-based classifiers could be effective due to the well-defined class regions. Given the strong lin-

ear characteristics, complex non-linear models are likely unnecessary.

These findings suggest that while the dataset presents some challenges (class imbalance, feature redundancy), its overall structure is favorable for classification tasks, particularly with linear models or simple neighborhood-based approaches.

IV. KEY CHALLENGES

A. Problexity Documentation Inconsistencies

The `T2` complexity measure (Average Number of Features per Dimension) presents documentation ambiguities regarding its terminology. While the formula uses "features over instances" in its calculation, there is confusion between features (dataset attributes) and dimensions (problem space representation). The correct name for this metrics is "Average Number of Features per Data Point". To address this ambiguity, I opened an issue in the `problexity` repository requesting clarification on the metric's documentation.

B. Computational Constraints

Neighborhood and network measures posed significant computational challenges due to the dataset size (70k records). These metrics, requiring pairwise distance calculations and graph-based structures, scale poorly with data volume. To address this limitation, a stratified sampling approach was implemented, reducing the dataset to 10% of its original size while preserving key characteristics for reliable complexity analysis. After applying stratified sampling, I compared several measures from the reduced dataset with those from the original dataset and the results were quite similar, indicating that the key characteristics of the dataset were preserved.

V. EXPERIMENTS AND TESTS ON ML ALGORITHMS

I tested some of the classical machine learning algorithms on the original dataset, using accuracy and F1 score as performance metrics. Given the high class imbalance, F1 score was prioritized to balance precision and recall. Hyperparameter tuning and cross-validation were employed to optimize algorithm performance.

A. Decision Trees

Decision Trees recursively partition the feature space to classify instances. We focused on the max_depth hyperparameter to prevent overfitting. The best performance was achieved with a max_depth of 5 (accuracy = 0.91 and F1 score = 0.952). While effective for this dataset, Decision Trees require careful tuning to avoid overfitting.

B. K-Nearest Neighbors (KNN)

KNN classifies based on the majority class of the nearest neighbors. We used Minkowski distance and normalized the dataset to improve performance. The key hyperparameter is the number of neighbors, with the best results at $k = 25$. KNN showed good performance with stable accuracy (0.91) and F1 score (0.953), but computational effort increases during prediction.

C. Logistic Regression

Logistic Regression is a linear model that estimates the probability of class membership. The regularization strength, 'C', was varied, but performance remained stable. The model achieved an accuracy of 0.908 and an F1 score of 0.952, confirming the effectiveness of the dataset's features.

D. Comparison of Algorithms

Model	Accuracy	F1	Time (s)
DecisionTree	0.909078	0.952374	0.113034
KNN	0.907877	0.951689	6.086654
LogisticRegression	0.908937	0.952296	0.026218

TABLE I. Model Performance Metrics

All three algorithms performed similarly, with accuracy around 0.91 and F1 score near 0.95. This consistency aligns with the dataset's strong linear separability, clear class boundaries and well-clustered instances in the data.

VI. REFERENCES

- Lorena, A. C., Garcia, L. P., Lehmann, J., Souto, M. C., & Ho, T. K. (2019). How complex is your classification problem? A survey on measuring classification complexity. *ACM Computing Surveys (CSUR)*, **52**(5), 1–34. ACM.
- Garcia, L. P., Lorena, A. C., de Souto, M. C., & Ho, T. K. (2018). Classifier recommendation using data complexity measures. In *2018 24th International Conference on Pattern Recognition (ICPR)* (pp. 874–879). IEEE.