Continuous-categorical feature dependence with Determination-based correlation coefficient*

Andrii Berko[1,†], Vladyslav Holdovanskyi[1,*,†]

[1]Lviv Polytechnic National University, 12 Stepana Bandery Street, Lviv, 79014, Ukraine

Abstract

The paper explores the application of the Determination-Based Correlation Coefficient (DBCC) for analyzing dependencies between continuous and categorical features in classification tasks. While DBCC has previously been validated in regression contexts, its suitability for classification problems remains underexplored. This study evaluates DBCC across various dependency structures, including step, multi-class, and sigmoidal relationships, demonstrating its robustness and effectivity. The findings suggest that DBCC provides a computationally efficient and flexible alternative to traditional methods such as ANOVA, point-biserial correlation, and mutual information. Additionally, classification performance metrics such as Accuracy, F1-score, and ROC-AUC are examined to assess the practical implications of using Grid-Mean Algorithm in feature selection and model evaluation. Future research directions include refining the methodology and extending its application to high-dimensional datasets.

Keywords

Analysis, Correlation coefficient, Classification, Categorical data, Machine Learning

1. Introduction

Understanding the relationship between continuous and categorical features is a fundamental aspect of many classification tasks in machine learning. While numerous techniques exist for dependency analysis, their effectiveness varies based on data structure, underlying assumptions, and computational efficiency. In our previous work, we introduced and analyzed the Determination-Based Correlation Coefficient (DBCC) in the context of regression tasks, demonstrating its ability to capture complex dependencies and maintain robust performance across diverse datasets [1]. However, classification problems require different approaches, as feature relationships often exhibit step or sigmoidal patterns that traditional correlation metrics struggle to interpret [2, 3, 4, 5, 6, 7]. Given these strengths, a natural question arises: can DBCC also be applied to classification problems, where feature relationships often exhibit stepwise or sigmoidal patterns that traditional correlation metrics struggle to interpret?

In this study, we explore the use of DBCC for assessing feature dependence in classification tasks, aiming to provide a robust and interpretable alternative to existing methods. We evaluate DBCC on synthetic datasets with stepwise, multi-class, and sigmoidal dependencies, comparing its performance against established dependency measures. Additionally, we examine classification metrics such as accuracy, F1-score, log loss, and ROC-AUC [8, 9, 10, 11, 12, 13] to assess how DBCC contributes to feature selection and model evaluation.

Accurate dependency estimation is crucial for classification models. Recent studies have highlighted the importance of effective dependency estimation in classification. Techniques such as ANOVA and point-biserial correlation rely on distributional assumptions that limit their applicability in real-world scenarios [14, 15, 16]. Additionally, mutual information (MI) provides a flexible alternative but often requires careful parameter selection and exhibits higher computational complexity [17, 18]. The Akaike Information Criterion (AIC) has been utilized in model selection but does not directly quantify feature dependence [19, 20, 21, 22]. Given these limitations, exploring alternative approaches like DBCC is

MoMLeT-2025: 7th International Workshop on Modern Machine Learning Technologies, June, 14, 2025, Lviv-Shatsk, Ukraine

*Corresponding author.

†These authors contributed equally.

andrii.y.berko@lpnu.ua (A. Berko); vladyslav.a.holdovanskyi@lpnu.ua (V. Holdovanskyi)

0000-0001-6756-5661 (A. Berko); 0009-0001-5751-0760 (V. Holdovanskyi)

essential for improving classification model performance.

2. Materials

In our previous studies, we conducted an in-depth investigation of the Determination-Based Correlation Coefficient (DBCC) and explored its application for predicting regression metrics based on data characteristics. Our findings demonstrated that DBCC is highly effective at capturing nonlinear dependencies, making it a powerful tool for assessing complex relationships within datasets. Given that classification tasks frequently involve interactions between continuous and categorical features, which often follow stepwise or sigmoidal patterns, we hypothesize that DBCC can also provide valuable insights in this context. DBCC is calculated using the following formula:

$$\eta^2 = 1 - \frac{\sum_{i=0}^{n} [y_i - M_x(y_i)]^2}{\sum_{i=0}^{n} [y_i - M(y)]^2}$$

where $M_x(y_i)$ – is the mean value within the grid cell that contains x_i. This grid-based mean is defined as follows: Let the interval $[a, b]$ be divided into N equally sized subintervals of width $h = (b - a)/N$. Then for each $x_i \in [a, b]$, the grid is:

$$\omega_h = \left\{ M_{x \in [x_j, x_{j+1}]}(y), \quad x_j = a + jh, \quad j = 0, \ldots, N, \quad h = \frac{b-a}{N} \right\}$$

To rigorously assess the effectiveness of DBCC in classification tasks, we designed and executed a series of controlled experiments using synthetic datasets. These experiments aimed to analyze how well DBCC captures different types of dependencies between continuous and categorical features. Our methodology consisted of the following key steps:

- Generating datasets with different types of dependencies between continuous and categorical features.
- Computing the DBCC correlation coefficient and interpreting its results.
- Analyzing the outcomes and assessing the adequacy of DBCC in various scenarios.

We categorized dependencies into five primary types to ensure comprehensive testing:

- Step function.
- Twice-changing step function.
- No dependency.
- Multi-class categorical feature.
- Sigmoidal dependency with varying levels of noise.

Each synthetic dataset contained 10,000 data points, ensuring stable and statistically meaningful estimates. To manage the discretization of continuous variables, we set the grid parameter to n = 100. This parameter controls the level of granularity in the approximation process: lower values smooth out dependencies, whereas higher values increase noise and inconsistencies. Through extensive testing, we determined that n = 100 provides an optimal balance between approximation accuracy and computational efficiency.

Step Function. In the first experiment, the continuous variable was sampled from a normal distribution with a mean of 50 and a standard deviation of 15. The categorical variable was assigned a value of 1 if the condition $X > 50$ was met and 0 otherwise (see Figure 1). This dependency structure is frequently observed in classification tasks where a clear threshold separates the classes.

The computed DBCC value was 0.975, indicating an almost perfect correspondence between the continuous feature and the categorical label. The approximation accurately captured the underlying

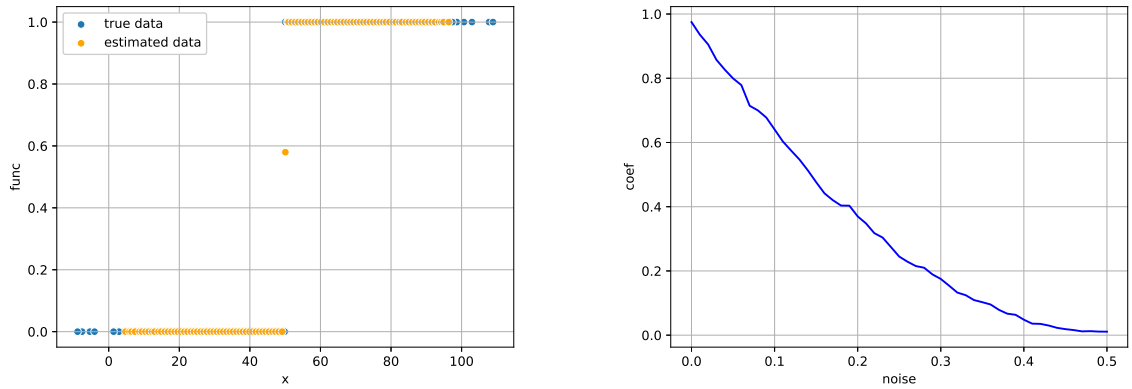

Figure 1: Illustration of Grid-based Algorithm approximation (left); DBCC under different noise levels for a double step function(right)

dependency, though minor uncertainty was observed around the threshold due to the effects of grid-based discretization.

Double Step Function. For the second experiment, we considered a more complex twice-changing step function. In this scenario, the categorical feature changed its value at two distinct thresholds, simulating real-world classification problems where decision boundaries are non-monotonic (see Figure 2).

The resulting DBCC value was 0.996, demonstrating DBCC's ability to effectively capture non-monotonic relationships. This result is particularly noteworthy, as many traditional correlation measures struggle to detect such patterns.

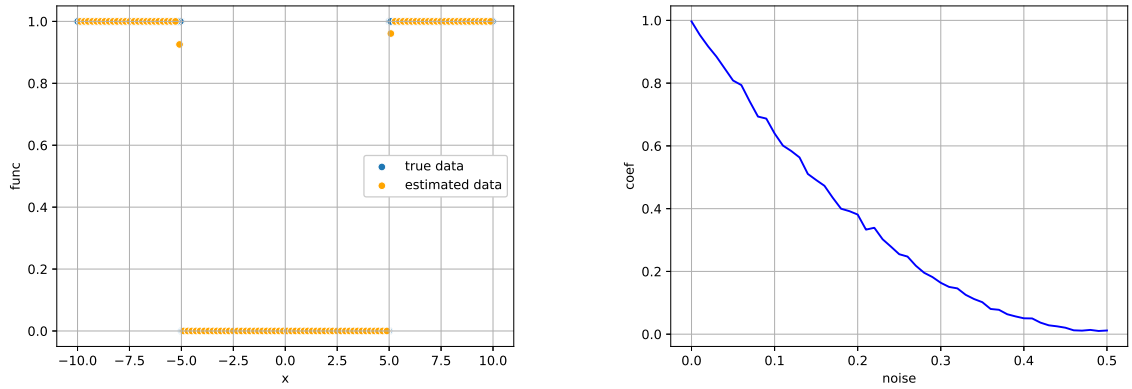

Figure 2: Illustration of Grid-based Algorithm approximation (left); DBCC under different noise levels for a step function(right)

No dependency. To examine DBCC's behavior when no dependency exists between the variables, we generated independent random features (see Figure 3). As expected, DBCC returned a near-zero correlation value of 0.009, correctly indicating the absence of a relationship. This outcome confirms that DBCC does not falsely identify correlations where none exist, reinforcing its reliability as a statistical measure.

Multiclass Categorical Feature. A distinguishing characteristic of DBCC is its capacity to handle multi-class categorical variables. To test this capability, we constructed a dataset where the categorical feature could assume three distinct values (see Figure 4). The analysis yielded a DBCC value of 0.984, illustrating DBCC's robustness in capturing relationships in multi-class classification settings.

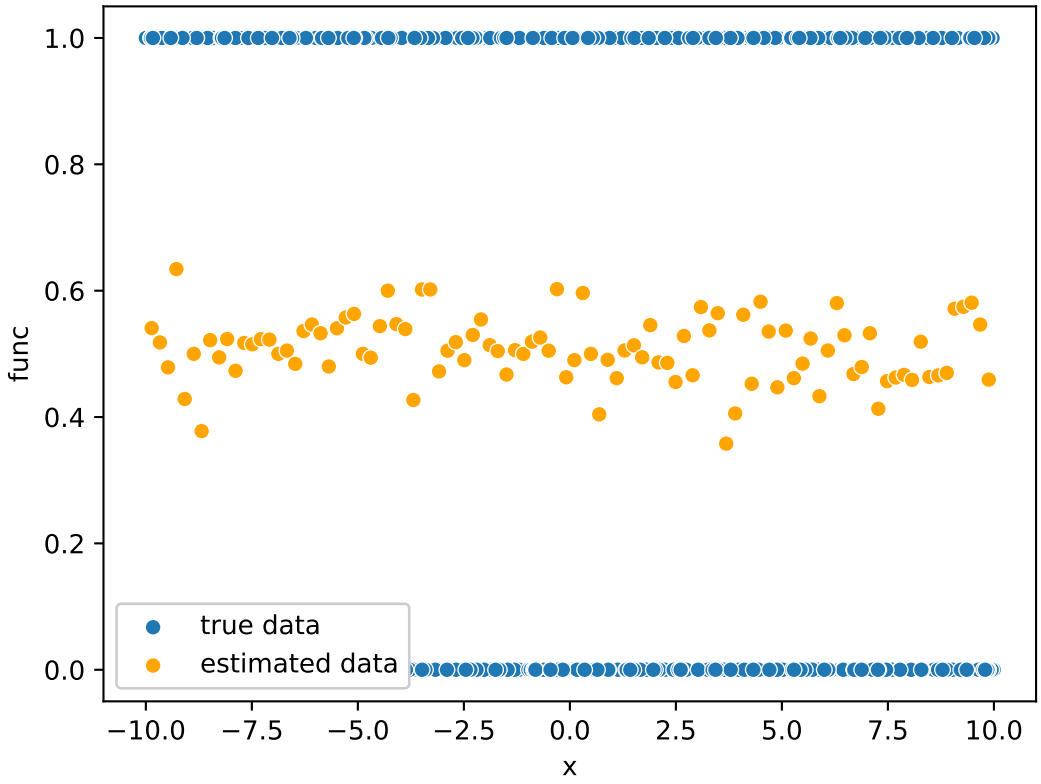

Figure 3: Illustration of Grid-based Algorithm approximation for a no dependency.

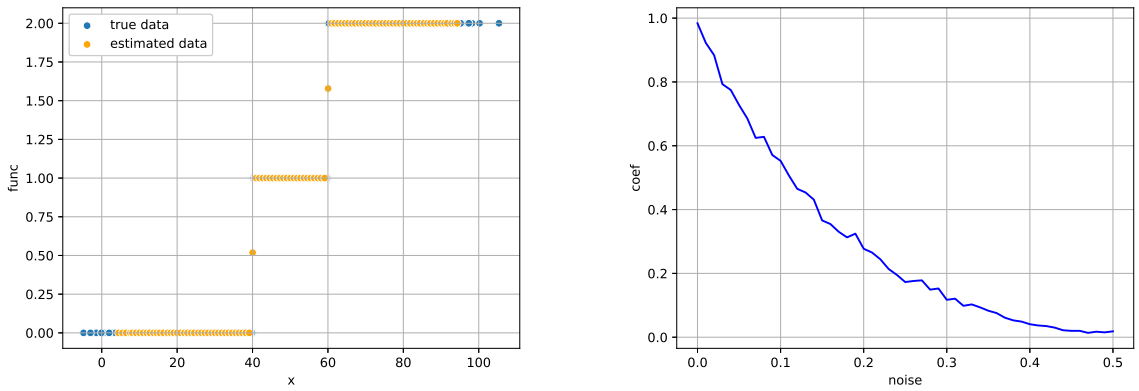

Figure 4: Illustration of Grid-based Algorithm approximation (left); DBCC under different noise levels for a multiclass categorical feature (right)

Sigmoidal Dependence. To simulate more intricate nonlinear dependencies, we introduced a sigmoidal function:

$$Y = \frac{1}{1 + e^{-\frac{X-50}{\lambda}}}$$

where λ controls the steepness of the transition between class labels. As λ increases, the function becomes more gradual, introducing greater uncertainty and blurring class boundaries (see Figure 5). Conversely, as λ approaches zero, the function approximates a power-law relationship.

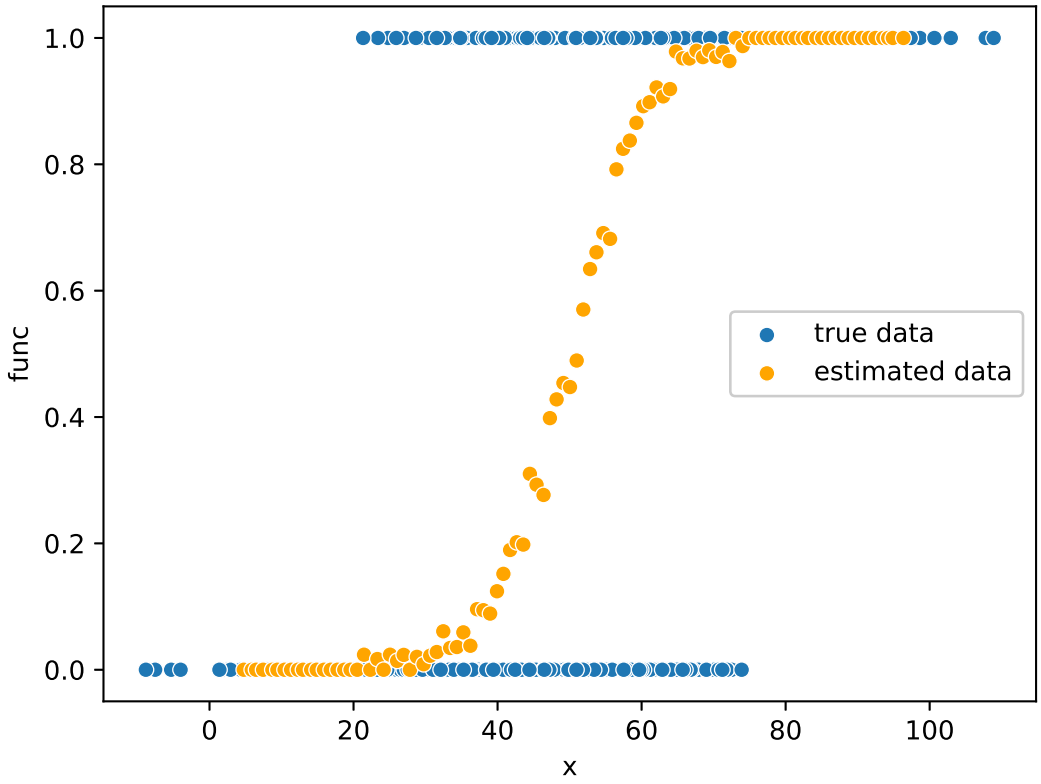

Figure 5: Illustration of Grid-based Algorithm approximation for a sigmoidal function.

Our results showed that as λ increased, the DBCC value decreased accordingly, reflecting the growing ambiguity in classification. This behavior suggests that DBCC is sensitive to uncertainty in decision boundaries and can serve as an effective diagnostic tool for evaluating the separability of classes in noisy data.

DBCC results showed that as class boundaries became more blurred (higher λ), the correlation coefficient decreased, reflecting the growing uncertainty in classification.

The findings indicate that DBCC is an effective tool for assessing dependencies in classification tasks. It demonstrates high accuracy even in cases of complex and nonlinear dependencies. In future research, we plan to evaluate the effectiveness of this approach on real-world datasets.

Comparison with Other Methods. To provide a broader perspective, we compared DBCC with widely used methods for analyzing dependencies between continuous and categorical variables. The key characteristics and performance metrics of these methods are summarized in Table 1.

Our preliminary results suggest that DBCC offers a unique balance between sensitivity to nonlinear patterns and robustness against noise, making it a valuable addition to the toolkit of data scientists working with classification problems.

Table 1
Comparison of different correlation methods

Method	Complexity	Params.	Distribution assumption	Interpretability	Output in [0, 1]	Vector Input
ANOVA/MANOVA	O(n)	-	Requires normal distribution	High	-	+
Kruskal-Wallis test	O(nlogn)	-	Same distribution in all groups	High	-	-
Point-biserial correlation	O(n)	-	Only two-category input	Medium	[-1;1]	-
Mutual Information	O(nlogn)	Num. of bins	No assumption	Medium	+	+
Akaike Criterion	O(n) + complexity of model	Model	Depends on model assumptions	High	[0, inf]	+
DBCC	O(n)	Num. of bins	Requires single-valued dependency	High	+	+

DBCC has proven to be an effective option, combining low computational complexity with good interpretability and the ability to handle multiclass categorical variables.

Key advantages of DBCC over other methods:

- **Computation speed:** Unlike MI, DBCC has linear computational complexity O(n), making it suitable for large datasets.
- **No strict distributional assumptions:** Unlike ANOVA, DBCC does not require normality of residuals or equal variances across groups.
- **Support for multiclass categorical variables:** Point-biserial correlation works only with binary variables, whereas DBCC can analyze multiple categories.
- **Interpretability:** The DBCC coefficient takes values in the range [0, 1,] allowing for an easy assessment of the strength of the relationship between variables.

Extending Grid-Mean Algorithm for Classification Metrics. In our other study, we explored the application of a grid-based algorithm for the approximate computation of regression metrics. The fundamental idea behind this approach was that by discretizing the continuous feature space into grid intervals and computing mean values within each interval, we could efficiently approximate regression-based evaluation metrics. Following the same logic, we propose extending this methodology to classification metrics, leveraging the Grid-Mean Algorithm as a machine learning model capable of estimating classification performance indicators.

Beyond correlation analysis, the grid-based structure of DBCC allows for the approximation of classification metrics by utilizing the probability distribution of categorical labels within each grid interval. Since each segment of the grid contains averaged feature values, it can serve as a foundation for estimating the probability of categorical outcomes, thereby supporting class prediction. This capability enables us to derive classification performance metrics directly from the grid without requiring an explicit predictive model.

We considered four primary classification metrics that are widely used for evaluating model performance:

- Accuracy – The proportion of correctly classified values relative to the total number of samples. This is a fundamental metric that provides an overall measure of predictive performance.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}}$$

- F1-Score – The harmonic mean of precision and recall, balancing the trade-off between false positives and false negatives. It is particularly useful in cases with imbalanced class distributions.

$$\text{F1-Score} = \frac{2 * \text{Precision} + \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

- ROC-AUC – The area under the receiver operating characteristic curve, which quantifies a model's ability to distinguish between classes across various threshold settings.
- Log Loss – The average logarithmic loss of predicted probabilities, where lower values indicate a better probabilistic prediction. This metric is particularly relevant for evaluating probabilistic classifiers.

$$\text{Log Loss} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \ln(p_i) + (1 - y_i)\ln(1 - p_i)\right]$$

where – p_i predicted probability of the i-point, and – y_i label of this point.

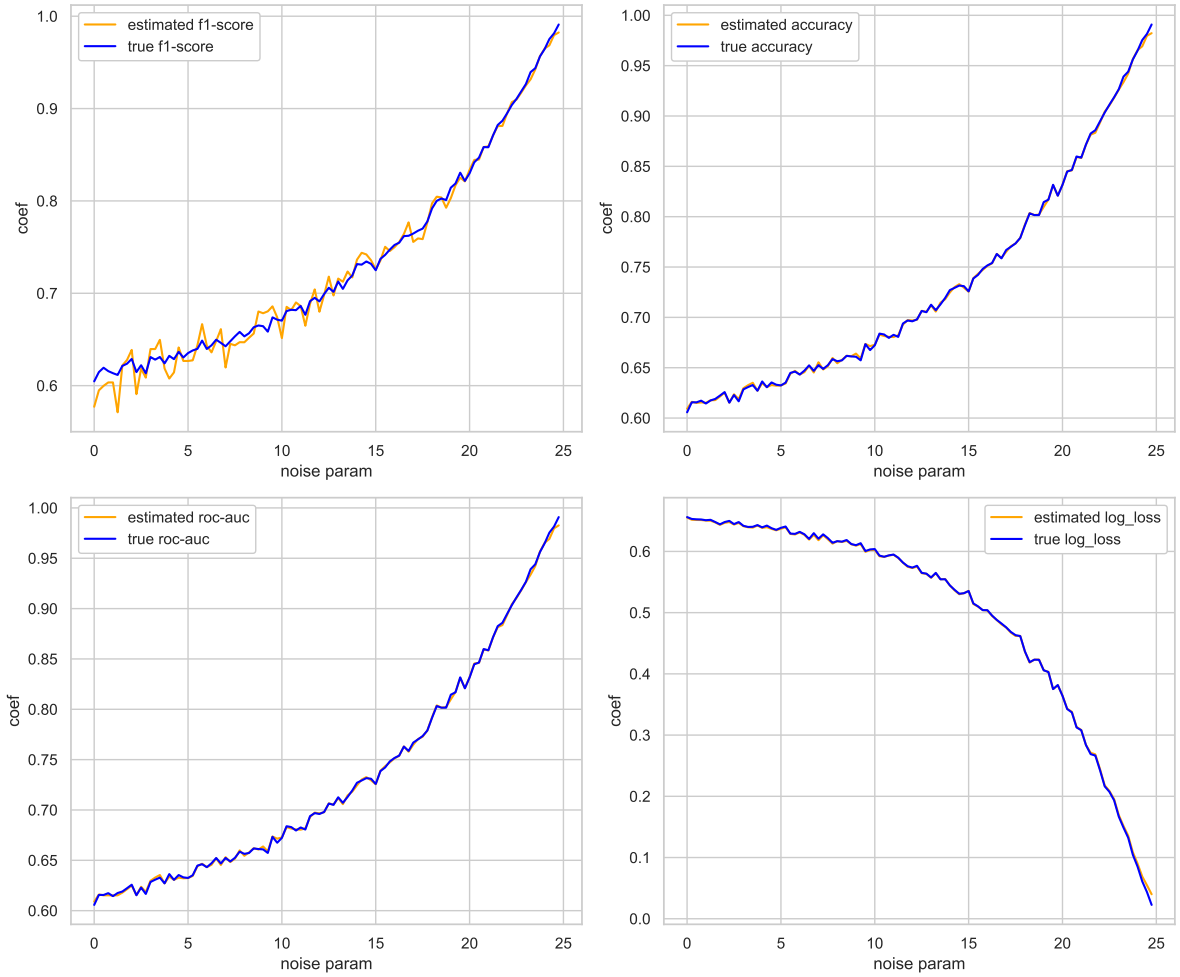

Figure 6: Comparison of true and estimated classification metrics values at different noise levels

To evaluate the effectiveness of this approach, we conducted experiments using a sigmoidal function with varying λ values:

$$Y = \frac{1}{1 + e^{-\frac{X-50}{\lambda}}}$$

where controls the level of noise in the data. Higher values of result in smoother transitions and greater uncertainty near the decision boundary, while lower values produce sharper class separations. This setup allows us to assess how well the Grid-Mean Algorithm can approximate classification metrics under different levels of ambiguity in class labels.

For each experimental setting, we compared the actual and approximated values of classification metrics, ensuring a comprehensive evaluation of this method's accuracy. The results (see Figure 6), demonstrate that the approximation error never exceeded 0.05 across all considered scenarios. This level of accuracy indicates that the proposed approach provides a reliable and computationally efficient means of estimating classification performance without the need for full-fledged model training.

Our findings indicate that DBCC is an effective tool for assessing dependencies in classification tasks. It demonstrates high accuracy, even in cases of complex and nonlinear dependencies. Moreover, its grid-based approach extends to approximate classification metric computation, making it a versatile tool in data analysis.

3. Conclusions

This study demonstrates the effectiveness of the Determination-Based Correlation Coefficient (DBCC) in evaluating dependencies between continuous and categorical features in classification tasks. Accurately assessing such dependencies is essential for improving classification models, as it influences feature selection, model interpretability, and predictive performance. Traditional methods often struggle to capture complex relationships, making the development of more flexible and efficient dependency measures a crucial research direction.

Through extensive experimentation on synthetic datasets, our results confirm that DBCC effectively identifies various types of relationships, including stepwise, sigmoidal, and multi-class dependencies, while also correctly recognizing the absence of correlation where applicable. The ability to handle both simple and intricate dependency structures makes DBCC a valuable tool for feature evaluation. Furthermore, compared to conventional approaches such as ANOVA, point-biserial correlation, and mutual information, DBCC offers a favorable balance between computational efficiency, interpretability, and adaptability to diverse data distributions.

Key Advantages of DBCC:

- Effective Handling of Nonlinear and Complex Dependencies: Unlike traditional methods that often assume linearity or specific distributions, DBCC can capture intricate relationships without imposing restrictive assumptions. This makes it particularly useful in real-world classification tasks, where dependencies are rarely straightforward.
- Robustness Across Various Dependency Structures: DBCC performs well across different types of relationships, including stepwise transitions, non-monotonic dependencies, and multi-class interactions. This generalizability enhances its applicability in diverse classification scenarios.
- High Computational Efficiency: One of DBCC's strengths is its ability to process large datasets efficiently, making it scalable for applications involving extensive feature analysis. Unlike computationally expensive techniques such as mutual information, DBCC provides a streamlined approach that balances precision with speed.

Future Research Directions. To further enhance DBCC's effectiveness, several avenues for future research should be explored. One key area is refining the methodology to improve precision and robustness across more diverse data structures. Advanced techniques for parameter tuning and optimization could enhance its sensitivity in detecting subtle dependencies, ensuring more reliable results even in challenging classification settings.

Another important direction is extending DBCC's applicability to high-dimensional datasets, where feature interactions become increasingly complex. Many real-world classification tasks involve numerous variables, and a robust dependency measure must be capable of efficiently handling such multi-feature relationships. Developing scalable implementations of DBCC that leverage parallel computing and optimized algorithms could significantly enhance its usability in large-scale machine learning workflows.

Moreover, integrating DBCC into automated feature selection frameworks could greatly benefit the field by providing a systematic way to identify relevant features based on their dependency structures. By incorporating DBCC into feature engineering pipelines, practitioners can streamline the process of selecting informative variables, ultimately improving model performance while reducing computational overhead.

Finally, exploring the potential of DBCC for approximate computation of classification metrics presents an exciting opportunity. If DBCC can serve as a proxy for evaluating classification performance, it could enable rapid model assessment without requiring extensive training and validation steps. Such an approach would be particularly valuable in iterative machine learning workflows, where quick feedback is essential for model optimization.

By addressing these research directions, DBCC can be further established as a powerful and versatile tool for analyzing complex dependencies in modern data science, contributing to more effective and interpretable classification models.

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI GPT-4 in order to: check grammar and spelling, improve formal writing style, and enhance overall readability of the text. After using this tool, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] A. Berko, V. Alieksieiev, V. Holdovanskyi, Determination-based correlation coefficient, in: Proceedings of the 6th International Workshop on Modern Machine Learning Technologies (MoMLeT-2024), Lviv-Shatsk, Ukraine, 2024. May 31 – June 1, 2024.

[2] S. B. Kotsiantis, I. D. Zaharakis, P. E. Pintelas, Machine learning: a review of classification and combining techniques, Artificial Intelligence Review (2007). doi:10.1007/s10462-007-9052-3.

[3] P. Pukach, B. Liubinskyy, V. Hladun, V. Holdovanskyi, The classifier models usage for the recruitment process forecasting for applicants of higher education to universities of ukraine, in: Data-Centric Business and Applications, Springer, 2024. doi:10.1007/978-3-031-60815-5_4.

[4] Q. An, S. Rahman, J. Zhou, J. J. Kang, A comprehensive review on machine learning in healthcare industry: Classification, restrictions, opportunities and challenges, Sensors 23 (2023). doi:10.3390/s23094178.

[5] A. F. A. H. Alnuaimi, T. H. K. Albaldawi, An overview of machine learning classification techniques, BIO Web of Conferences 97 (2024) 00133. doi:10.1051/bioconf/20249700133.

[6] R. I. Mukhamediev, Y. Popova, Y. Kuchin, et al., Review of artificial intelligence and machine learning technologies: Classification, restrictions, opportunities and challenges, Mathematics 10 (2022) 2552. doi:10.3390/math10152552.

[7] J. E. Black, J. K. Kueper, T. S. Williamson, An introduction to machine learning for classification and prediction, Family Practice (2022). doi:10.1093/fampra/cmac104.

[8] L. Ferrer, Analysis and comparison of classification metrics, arXiv preprint arXiv:2209.05355, 2023.

[9] Z. D. Vujović, Classification model evaluation metrics, International Journal of Advanced Computer Science and Applications 12 (2021). doi:10.14569/IJACSA.2021.0120670.

[10] J. Zhou, A. H. Gandomi, F. Chen, A. Holzinger, Evaluating the quality of machine learning explanations: A survey on methods and metrics, Electronics 10 (2021) 593. doi:10.3390/electronics10050593.

[11] M. Steurer, R. J. Hill, N. Pfeifer, Metrics for evaluating the performance of machine learning-based automated valuation models, Journal of Property Research 38 (2021) 99–129. doi:10.1080/09599916.2020.1858937.

[12] J. Chen, Analysis of statistic metrics in different types of machine learning, in: Highlights in Science, Engineering and Technology IFMPT, volume 88, 2024, p. 182.

[13] S. A. Hicks, I. Strümke, V. Thambawita, et al., On evaluation metrics for medical applications of artificial intelligence, Scientific Reports 12 (2022) 5979. doi:10.1038/s41598-022-09954-8.

[14] C. E. Flores, K. L. Flores, Kruskal-wallis, friedman and mood nonparametric tests applied to business decision making, Espirales. Revista Multidisciplinaria de Investigación Científica 6 (2022) 1–20.

[15] D. C. Howell, Fundamental Statistics for the Behavioral Sciences, 5th ed., Duxbury Press, Pacific Grove, 2004.

[16] D. J. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, 2nd ed., Chapman & Hall, London, 2000.

[17] A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information, Physical Review E 69 (2004) 066138. doi:10.1103/PhysRevE.69.066138.

[18] N. Veyrat-Charvillon, F.-X. Standaert, Mutual information analysis: How, when and why?, in:

Cryptographic Hardware and Embedded Systems - CHES 2009, volume 5747 of *Lecture Notes in Computer Science*, Springer, 2009, pp. 429–443.

[19] Q. Liu, M. A. Charleston, S. A. Richards, B. R. Holland, Performance of akaike information criterion and bayesian information criterion in selecting partition models and mixture models, Systematic Biology 72 (2023) 92–105. doi:10.1093/sysbio/syac081.

[20] J. E. Cavanaugh, A. A. Neath, The akaike information criterion: Background, derivation, properties, application, interpretation, and refinements, Wiley Interdisciplinary Reviews: Computational Statistics 11 (2019) e1460. doi:10.1002/wics.1460.

[21] H. Cheng, B. Sterner, Error statistics using the akaike and bayesian information criteria, Erkenntnis (2024). doi:10.1007/s10670-024-00897-2.

[22] C. Sutherland, D. Hare, P. J. Johnson, et al., Practical advice on variable selection and reporting using akaike information criterion, Proceedings of the Royal Society B: Biological Sciences 290 (2023) 20231261. doi:10.1098/rspb.2023.1261.