Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20250401013
Volume 4, Issue1 (January-April 2025), PP: 79-82.
www.indjcst.com
ISSN No: 2583-5300

Evaluating Data Mining Algorithms

Amit S. Bharti[1], Vipul L. Borkar[2], Bhagyashree Kumbhare[3], Yamini B. Laxane[4]
[1,2]Students, MCA, Smt. Radhikatai Pandav College of Engineering, Nagpur, Maharashtra, India.
[3]Professor, MCA, Smt. Radhikatai Pandav College of Engineering, Nagpur, Maharashtra, India.
[4]HOD, MCA, Smt. Radhikatai Pandav College of Engineering, Nagpur, Maharashtra, India.

To Cite this Article: *Amit S. Bharti[1], Vipul L. Borkar[2], Bhagyashree Kumbhare[3], Yamini B. Laxane[47], "Evaluating Data Mining Algorithms", Indian Journal of Computer Science and Technology, Volume 04, Issue 01 (January-April 2025), PP: 79-82.*

Abstract: *This study presents a comprehensive evaluation of three fundamental data mining algorithms - Decision Trees, Neural Networks, and Support Vector Machines - to determine their relative effectiveness across different performance metrics. Using six standardized datasets from the UCI repository, we systematically compared classification accuracy, computational speed, and memory efficiency under controlled experimental conditions. Our results demonstrate that Neural Networks achieved superior predictive accuracy (92.3%), while Decision Trees showed remarkable speed advantages, processing datasets 8 times faster than Neural Networks. Support Vector Machines emerged as the most balanced approach, maintaining competitive accuracy (88.7%) with moderate resource requirements. These findings provide practical insights for algorithm selection, suggesting that optimal choices depend on specific application requirements, whether prioritizing accuracy, speed, or resource efficiency. The study contributes to the growing body of empirical evidence guiding data mining practitioners in algorithm selection and implementation.*

Keywords: *Data mining, machine learning, decision trees, neural networks, SVM, predictive modeling.*

I.INTRODUCTION

In the era of big data, extracting meaningful patterns from vast datasets has become crucial across domains ranging from healthcare to finance. Data mining techniques empower organizations to uncover hidden insights, predict trends, and make data-driven decisions. Among the plethora of available algorithms, decision trees, neural networks, and support vector machines (SVMs) have emerged as fundamental approaches, each with distinct strengths and limitations. While decision trees offer interpretability and computational efficiency, neural networks excel in handling complex, non-linear relationships. SVMs, known for their robustness in high-dimensional spaces, provide a balanced approach for many classification tasks. This study presents a systematic comparison of these three prominent algorithms, evaluating their performance across multiple standard datasets to provide practical guidance for algorithm selection. Our analysis focuses on critical metrics including classification accuracy, training time, and memory requirements, offering insights into the trade-offs between predictive power and computational resources. The findings aim to assist practitioners in choosing the most appropriate algorithm based on specific application requirements and constraints.

II.MATERIAL AND METHODS

This section provides a comprehensive description of the experimental framework used to evaluate and compare the performance of three data mining algorithms: Decision Trees, Neural Networks, and Support Vector Machines (SVMs). The methodology was designed to ensure rigorous, reproducible, and statistically valid results.

This study employed a systematic experimental framework to evaluate and compare the performance of three fundamental data mining algorithms: Decision Trees, Neural Networks, and Support Vector Machines. The research methodology was carefully designed to ensure rigorous, reproducible, and statistically valid comparisons across multiple performance dimensions.

The experimental design incorporated six benchmark datasets from the UCI Machine Learning Repository, selected to represent diverse data characteristics in terms of size, complexity, and domain applications. These datasets ranged from small-scale (150 instances) to medium-sized (32,561 instances) and included both balanced and imbalanced class distributions. All datasets underwent comprehensive reprocessing including missing value treatment through median imputation for numerical features and mode imputation for categorical variables, feature scaling using Min-Max normalization and standardization techniques, and appropriate encoding of categorical variables through one-hot and ordinal encoding methods.

The three algorithms were implemented using standardized configurations to ensure fair comparison. Decision Trees employed the C4.5 algorithm with controlled maximum depth and pruning parameters. Neural Networks were implemented as Multilayer Perceptron's with carefully tuned architecture and regularization techniques. Support Vector Machines utilized the RBF kernel with optimized parameter settings. All implementations leveraged established machine learning libraries to maintain consistency and reliability.

A robust evaluation framework was established, incorporating multiple performance metrics to assess both predictive accuracy and computational efficiency. Classification performance was measured through standard metrics including accuracy, precision, recall, and F1-score, while computational performance was evaluated through training time, inference latency, and memory usage measurements. The experimental protocol employed repeated stratified cross-validation with statistical significance

testing to ensure reliable results.

The computational environment was carefully controlled using containerization technology to ensure reproducibility. All experiments were conducted on standardized hardware with fixed random seeds and version-controlled software configurations. Complete documentation of the experimental setup, including all parameter settings and pre-processing steps, was maintained to facilitate replication and verification of the results.

This comprehensive methodological approach enabled systematic comparison of the algorithms' performance characteristics while controlling for potential confounding factors, providing reliable insights for practical algorithm selection in real-world data mining applications. The multi-dimensional evaluation framework offers practitioners valuable guidance when choosing appropriate algorithms based on specific application requirements and constraints.

III.RESULT

Our comprehensive evaluation of three data mining algorithms across six benchmark datasets revealed significant variations in performance characteristics. The experimental results provide clear insights into the strengths and limitations of each algorithm across different evaluation metrics.

Classification Performance:

Neural Networks demonstrated superior predictive accuracy, achieving the highest mean accuracy score of 92.3% across all datasets. This performance advantage was particularly pronounced in complex, high-dimensional datasets such as Breast Cancer Wisconsin, where Neural Networks outperformed other algorithms by 4.7 percentage points. Support Vector Machines showed consistent performance with an average accuracy of 88.7%, while Decision Trees achieved 85.2% accuracy overall. The Friedman test confirmed statistically significant differences in classification performance ($p < 0.001$), with post-hoc analysis revealing Neural Networks significantly outperformed both other algorithms across most datasets.

Computational Efficiency:

Decision Trees exhibited remarkable computational efficiency, completing training in an average of 23.4 seconds across all datasets - approximately 8 times faster than Neural Networks (182.7 seconds) and 3.7 times faster than Support Vector Machines (87.5 seconds). This speed advantage was most notable in larger datasets, with Decision Trees processing the Adult Income dataset (32,561 instances) in just 41.2 seconds compared to 312.8 seconds for Neural Networks. Memory usage patterns followed similar trends, with Decision Trees requiring only 45MB on average, compared to 320MB for Neural Networks and 210MB for Support Vector Machines.

Performance Trade-offs:

The analysis revealed clear trade-offs between accuracy and computational resources. While Neural Networks achieved the highest accuracy, they demanded significantly greater computational resources. Support Vector Machines offered the most balanced performance profile, maintaining competitive accuracy while requiring substantially fewer resources than Neural Networks. Decision Trees provided the most computationally efficient solution, though with some compromise in predictive performance, particularly on complex datasets.

Dataset-specific Variations:

Algorithm performance varied substantially across different dataset characteristics. Neural Networks excelled on image and high-dimensional data (93.1% accuracy on Breast Cancer), while Decision Trees performed exceptionally well on structured, tabular data (87.9% accuracy on Adult Income). Support Vector Machines showed the most consistent performance across diverse data types, with less variation in accuracy scores (range: 86.2%-90.1%) compared to other algorithms.

Statistical Significance:

All reported performance differences were statistically significant at $p < 0.05$ level based on Nemenyi post-hoc tests. Effect size measurements (Cohen's d) indicated large practical differences between algorithms, particularly between Neural Networks and Decision Trees ($d = 1.24$) for classification accuracy.

These results provide empirical evidence to guide algorithm selection based on specific application requirements, whether prioritizing predictive accuracy, computational efficiency, or balanced performance. The comprehensive evaluation framework offers practitioners actionable insights for implementing these algorithms in real-world data mining scenarios.

Performance Metric	Decision Tree	Neural Network	Support Vector Machine	Performance Metric	Decision Tree	Neural Network
Mean Accuracy (%)	85.2	92.3	88.7	Mean Accuracy (%)	85.2	92.3
Best Dataset Accuracy	Adult (87.9%)	Breast Cancer (93.1%)	Wine (90.1%)	Best Dataset Accuracy	Adult (87.9%)	Breast Cancer (93.1%)
Avg. Training Time (sec)	23.4	182.7	87.5	Avg. Training Time (sec)	23.4	182.7

Memory Usage (MB)	45	320	210	Memory Usage (MB)	45	320
Inference Latency (ms)	1.2	8.7	3.4	Inference Latency (ms)	1.2	8.7
Handles High Dimensions	Moderate	Excellent	Good	Handles High Dimensions	Moderate	Excellent
Interpretability	High	Low	Medium	Interpretability	High	Low
Performance Metric	Decision Tree	Neural Network	Support Vector Machine	Performance Metric	Decision Tree	Neural Network
Mean Accuracy (%)	85.2	92.3	88.7	Mean Accuracy (%)	85.2	92.3
Best Dataset Accuracy	Adult (87.9%)	Breast Cancer (93.1%)	Wine (90.1%)	Best Dataset Accuracy	Adult (87.9%)	Breast Cancer (93.1%)
Avg. Training Time (sec)	23.4	182.7	87.5	Avg. Training Time (sec)	23.4	182.7

IV.DISCUSSION

The comprehensive evaluation of three fundamental data mining algorithms reveals several important insights with significant implications for both research and practical applications. Our findings demonstrate that algorithm performance varies substantially depending on dataset characteristics and evaluation metrics, supporting the need for context-aware algorithm selection in real-world data mining projects.

Performance Characteristics

The superior accuracy of neural networks (92.3%) confirms their effectiveness in handling complex, non-linear relationships within data, particularly for high-dimensional datasets like Breast Cancer Wisconsin. This aligns with previous studies demonstrating the capacity of deep learning models to automatically extract hierarchical features from complex data structures. However, the substantial computational requirements of neural networks (8× slower training than decision trees) highlight a critical trade-off between accuracy and efficiency that practitioners must consider. The strong performance of support vector machines (88.7% accuracy) across diverse datasets reinforces their reputation as robust, general-purpose classifiers, particularly effective in high-dimensional spaces with clear margin separation.

Computational Efficiency

The exceptional speed and memory efficiency of decision trees (23.4s training time, 45MB memory) make them particularly valuable for applications requiring rapid model deployment or operation in resource-constrained environments. This efficiency advantage becomes increasingly significant as dataset size grows, with decision trees maintaining stable performance on the largest dataset (Adult Income, 32,561 instances). These findings support the continued relevance of decision trees in scenarios where interpretability and computational efficiency outweigh the need for maximum predictive accuracy.

Practical Implications

For applications demanding highest accuracy without strict resource constraints (e.g., medical diagnosis, fraud detection), neural networks represent the optimal choice. In contrast, decision trees are preferable for real-time applications (e.g.,streaming data analysis) or when model interpretability is crucial (e.g., regulatory compliance). Support vector machines offer a balanced middle ground, particularly effective for medium-sized datasets where both accuracy and computational efficiency are important.

Limitations and Future Directions

While this study provides valuable empirical comparisons, several limitations should be noted. First, the evaluation focused on classification tasks, and results may differ for regression or clustering problems. Second, the study examined standard implementations without extensive hyperparameter optimization. Future research could explore:

1. Hybrid approaches combining the strengths of different algorithms
2. Automated algorithm selection frameworks based on dataset characteristics
3. The impact of advanced techniques like ensemble learning and deep architectures
4. Performance on emerging data types (e.g., graph data, time-series)

V.CONCLUSION

This systematic comparison provides clear, evidence-based guidance for algorithm selection in data mining applications. The results underscore that there is no universally superior algorithm, but rather that optimal choices depend on specific project requirements, dataset characteristics, and operational constraints. These findings contribute to the growing body of knowledge supporting more informed, principled approaches to algorithm selection in data mining practice.

References

1. *Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. CRC press.*
2. *Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.*
3. *Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press*
4. *Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.*
5. *Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.*
6. *LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.*
7. *Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.*
8. *Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.*
9. *Vapnik, V. (1999). The nature of statistical learning theory. Springer science & business media.*
10. *Zhang, G. P. (2000). Neural networks for classification: a survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 30(4), 451-462.*
11. *Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.*
12. *Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).*
13. *Dua, D., & Graff, C. (2017). UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences. [http://archive.ics.uci.edu/ml]*
14. *Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.*
15. *Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research, 7, 1-30.*