Comparative Analysis of Bayesian Networks and Tree-Based Methods for Breast Cancer Classification

Sai Krishna Vavilli

23022047

Introduction

This study presents a comparative analysis of machine learning approaches applied to the Wisconsin Breast Cancer dataset, focusing on Bayesian Networks, Decision Trees, and Random Forests. The primary objective is to evaluate these methods' effectiveness in classifying breast cancer cases as benign or malignant, providing insights into their relative strengths and applications in medical diagnosis.

Dataset and Preprocessing

The Wisconsin Breast Cancer dataset comprises diagnostic measurements from digitized images of breast mass FNAs. The dataset contains 569 samples with 30 features, representing various cellular nucleus characteristics. Key preprocessing steps included:

- Data cleaning: No missing values were identified
- Feature standardization: Applied to normalize numerical features
- Train-test split: 80-20 ratio with stratification

Methodology

Bayesian Network

The implemented Bayesian Network model incorporated:

- Directed acyclic graph structure focusing on key feature dependencies
- Discretization of continuous variables into four categories
- MaximumLikelihoodEstimationforparameterlearning
- Variable Elimination for inference

Decision Tree

The Decision Tree classifier was implemented with:

- Maximum depth: 5 (preventing overfitting)
- Giniimpuritycriterion
- Controlled random state for reproducibility

Random Forest

The Random Forest ensemble

utilized: • 100 decision tree

estimators • Bootstrapsampling •

Feature randomization at splits

Results

Performance Metrics Comparison

Model	Accuracy	F1-Score
DecisionTree	0.947	0.958
RandomForest	0.965	0.972
BayesianNetwork	0.860	0.881

Model Characteristics

Each model exhibited distinct advantages:

- Random Forest: Optimal for pure prediction tasks
- Bayesian Network: Valuable for understanding feature relationships
- Decision Tree: Superior interpretability for stakeholder communication

Conclusion

This study demonstrates the effectiveness of both probabilistic and tree-based approaches in breast can- cer classification. While Random Forest achieved the highest accuracy, each method offers unique advantages for different application contexts.

References

2. Friedman, N., Geiger, D. & Goldszmidt, M. (1997) ’Bayesian Network Classifiers’, Machine

1. Street, W.N., Wolberg, W.H. & Mangasarian, O.L. (1993) ’Nuclear feature extraction for breast tumor diagnosis’, IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology. San Jose, CA, pp. 861-870.

Learning, 29(2), pp. 131-163.

3. Breiman, L. (2001) ’Random Forests’, Machine Learning, 45(1), pp. 5-32.

4. Quinlan, J.R. (1986) ’Induction of Decision Trees’, Machine Learning, 1(1), pp. 81-106.