WITS
UNIVERSITY

University of Witwatersrand

Probabilistic Graphical Models

Probabilistic Graphical Model for Diabetes Diagnosis

Research Report

Author:
Sergio Frasco
2427724
April 28, 2024

1 Introduction

Diabetes is a chronic metabolic disorder that affects millions of people worldwide, and its early detection and accurate diagnosis are crucial for effective management and prevention of complications. This research explores the application of Probabilistic Graphical Models (PGMs) in classifying diabetes using a dataset of health indicators. PGMs provide a powerful framework for representing and reasoning about uncertain relationships among variables, making them suitable for medical diagnosis tasks. The objective of this study is to develop a PGM-based algorithm that can accurately classify diabetes in patients based on their health indicators and to evaluate its performance using various metrics.

2 Related Work

Previous studies have applied various machine learning techniques for diabetes classification, including logistic regression, decision trees, and support vector machines [4]. However, the use of PGMs in this context has been limited. Existing research has demonstrated the potential of PGMs in medical diagnosis tasks, such as in the classification of breast cancer [3] and the prediction of Alzheimer's disease [1]. This study aims to extend the application of PGMs to diabetes classification and propose novel techniques for model optimization.

3 Dataset and Preprocessing

The dataset used in this study, obtained from the UCI Machine Learning Repository, consists of 47 different health indicators collected from patients as well as over 100,000 instances. The data preprocessing pipeline involves several key steps to ensure data quality and prepare the dataset for model learning.

1. Missing data handling: A Bayesian network approach is employed to estimate missing values based on the probabilistic relationships among variables. The Hill-Climb Search algorithm with the Bayesian Information Criterion (BIC) score is used to learn the network structure, and the Variable Elimination inference method is applied to fill in missing data points.

2. Binning: Numerical columns are discretized through binning to simplify the representation of continuous features, reduce model complexity, and improve generalzation by reducing the impact of outliers and noise.

3. Feature selection: The most informative features are identified based on their correlation with the target variable (diabetes). This step reduces dimensional-

ity, improves computational efficiency, and prevents overfitting by eliminating irrelevant or redundant features.

4. Oversampling: Techniques such as Random Oversampling (ROS) or Synthetic Minority Over-sampling Technique (SMOTE) are applied to address class imbalance. This ensures that the model learns from a balanced dataset and improves its ability to correctly classify both diabetic and non-diabetic patients.
5. Scaling: Min-Max scaling is used to normalize the features to a fixed range (0 to 1), preventing any single feature from dominating the learning process and ensuring that all features contribute equally to the model's predictions.
6. Splitting: The data is then split into training, validation and test data. This ratio is of the form 70:20:10.

These preprocessing steps enhance the quality of the data and set the stage for effective model learning and accurate diabetes classification using the Probabilistic Graphical Model (PGM) approach.

4 Methodology

4.1 PGM Selection and Structure Learning

A Bayesian network is selected as the PGM for diabetes classification due to its ability to represent probabilistic relationships among variables, perform inference based on observed evidence as well as providing an intuitive graphical structure for representing relationships. The structure of the Bayesian network is learned using the Hill-Climb Search algorithm with random restarts. The random restarts technique helps in exploring different network configurations and finding the optimal structure by avoiding local optima. The algorithm compares different scoring methods, including K2Score, BDeuScore and BICScore, to assess their impact on model performance. Following this, the best model is selected based on it's metrics obtained by comparing the random restart graphs as well as the various scoring methods. This model is then used to perform the inference tasks.

While Bayesian networks excel in capturing probabilistic dependencies among variables, they may struggle with modeling complex, non-linear relationships present in some datasets. Additionally, their performance can be hindered by the curse of dimensionality, especially when dealing with a large number of variables or high dimensional data. Despite these limitations, Bayesian networks were selected for diabetes classification due to their interpretability and suitability for modeling uncertainty in medical diagnosis tasks.

4.2 Inference and Prediction

Variable Elimination emerges as the method of choice for predicting diabetes from health indicators due to its notable reliability and speed. By efficiently processing probabilistic relationships among variables, Variable Elimination ensures robust predictions while maintaining computational efficiency. Its systematic approach of eliminating variables while preserving accuracy streamlines the prediction process, allowing for rapid evaluation of numerous scenarios. This reliability and speed are particularly advantageous in healthcare settings, where timely and precise predictions are essential for informed decision-making and patient care. Additionally, Variable Elimination 's ability to handle complex datasets and incorporate evidence in real-time enhances its practical utility in clinical applications. Overall, Variable Elimination stands as a reliable and fast-paced solution for forecasting diabetes, offering a blend of accuracy, efficiency, and versatility suited for healthcare contexts.

5 Results and Discussion

The achieved evaluation metrics (Accuracy: 76%, Precision: 75%, Sensitivity: 80%, F1-score: 77%) shed light on the characteristics of the diabetes data. The accuracy of 76% indicates the overall correctness of predictions, suggesting that the algorithm performs reasonably well in classifying both diabetic and non-diabetic patients. The precision score of 75% reflects the proportion of true positive predictions among all positive predictions, highlighting the algorithm's ability to avoid false positives. With a sensitivity of 80%, the algorithm demonstrates its capability to correctly identify diabetic patients without missing positive cases, crucial for early detection and intervention in diabetes management. The F1-score of 77% provides a balanced measure of the algorithm's performance, considering both precision and sensitivity [2]. These scores collectively suggest that while the algorithm shows promising results, there may still be room for improvement, particularly in minimizing false negatives and enhancing overall classification accuracy.

The confusion matrix, depicted in Figure 1, offers a comprehensive overview of the algorithm's diabetes classification performance. It visualizes true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Notably, the matrix reveals a substantial number of correctly identified diabetic and non-diabetic patients (TP and TN), with minimal false positives and false negatives. This demonstrates the algorithm's effectiveness in minimizing misclassifications. The clarity and intuitiveness of the confusion matrix enable healthcare professionals to evaluate the algorithm's reliability and make informed decisions based on its predictions.

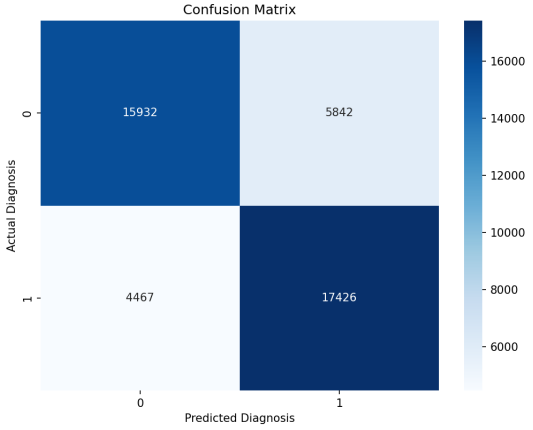

Figure 1: Confusion Matrix

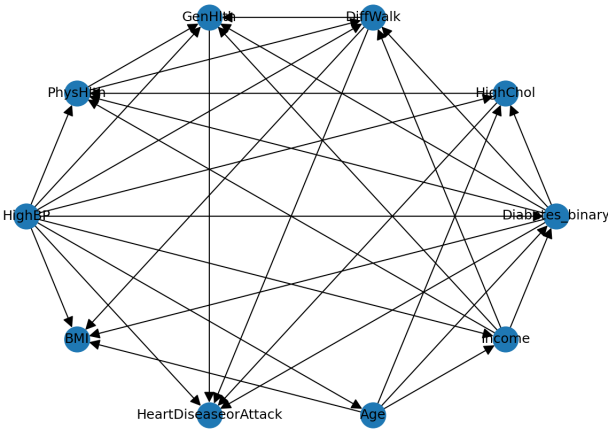

Figure 2: Confusion Matrix

The Bayesian network structure in Figure 2 captures the probabilistic relationships among health indicators and diabetes. It reveals a sparse connectivity pattern, indicating significant conditional dependencies within the dataset. This suggests a robust and comprehensive model. The utilization of random restarts and diverse scoring methods enhances the model's stability and adaptability. These techniques allow for identifying the optimal network structure and selecting suitable scoring functions. Furthermore, the algorithm's capability to handle missing data, conduct feature selection, and address class imbalance emphasizes its practicality in real world scenarios.

6 Limitations and Future Work

While the PGM algorithm shows promising results, it is important to acknowledge the limitations of this study on top of the limitations of bayesian networks we have already discussed in the relevant sections. The dataset used may not represent the full diversity of patient populations, and further validation on larger and more diverse datasets is necessary to assess the algorithm's generalizability. Additionally, the algorithm's performance should be compared with other state-of-the-art machine learning techniques to evaluate its relative effectiveness.

Future work could explore the integration of additional health indicators and the incorporation of temporal information to capture the progression of diabetes over time. Moreover, the interpretability of the learned Bayesian network could be investigated to provide insights into the relationships among health indicators and their impact on diabetes diagnosis. Collaborating with healthcare professionals and domain experts could further refine the algorithm and ensure its alignment with clinical practices.

7 Conclusion

This study showcases the application of Probabilistic Graphical Models (PGMs) for diabetes classification using health indicators. The proposed PGM algorithm, leveraging techniques such as data preprocessing, structure learning, and inference, yields promising results in accurately predicting diabetes. The evaluation metrics and confusion matrix emphasize the algorithm's effectiveness in identifying diabetic patients.

These findings contribute to the field by highlighting the potential of PGMs in aiding diabetes diagnosis and enhancing patient care. The algorithm's robustness to real-world datasets and diverse learning conditions positions it as a valuable tool for healthcare professionals. Nevertheless, further research is needed to validate its performance on larger and more varied datasets and to compare it with other state-of-the-art machine learning approaches. Collaborative efforts with healthcare experts can augment the algorithm's accuracy and practicality.

In summary, the presented PGM algorithm for diabetes classification demonstrates the effects of probabilistic reasoning in medical diagnosis. With ongoing research and refinement, such approaches hold promise for early detection, personalized treatment as well as improved patient outcomes in diabetes management.

References

[1] Athanasios Alexiou, Vasileios D Mantzavinos, Nigel H Greig, and Mohammad A Kamal. A bayesian model for the prediction and early diagnosis of alzheimer's disease. *Frontiers in aging neuroscience*, 9:77, 2017.

[2] Cyril Goutte and Eric Gaussier. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European conference on information retrieval*, pages 345–359. Springer, 2005.

[3] Mahmoud Khademi and Nedialko S Nedialkov. Probabilistic graphical models and deep belief networks for prognosis of breast cancer. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 727–732. IEEE, 2015.

[4] Bahman P Tabaei and William H Herman. A multivariate logistic regression equation to screen for diabetes: development and validation. *Diabetes Care*, 25(11):1999–2003, 2002.