Estudio sobre el ajuste óptimo de pesos en SPODEs para balancear *Accuracy* y *Fairness* en el clasificador probabilístico AODE

M. Julia Flores*
Departmento de Sistemas Informáticos - I3A
Universidad de Castilla - La Mancha
Julia.Flores@uclm.es

José A. Gámez
Departmento de Sistemas Informáticos - I3A
Universidad de Castilla - La Mancha
Jose.Gamez@uclm.es

Abstract

Este proyecto propone un enfoque basado en optimización multiobjetivo y algoritmos evolutivos para parametrizar el clasificador Average One-Dependence Estimators (AODE), mediante un sistema de pesado de los diferentes SPODEs. El estudio está centrado en el impacto que este pesado produce respecto a la mejora en cuanto a medidas de fairness (equidad) para este clasificador, sin menoscabar en exceso su precisión. Se utilizará Naïve Bayes como baseline para comparar los resultados. La optimización se llevará a cabo mediante técnicas evolutivas como NSGA-II o MOEA/D, buscando configuraciones óptimas que reduzcan sesgos en conjuntos de datos con atributos sensibles. Se evaluará el impacto del ajuste de pesos en varios datasets, utilizando métricas estándar de clasificación y equidad. Finalmente, se analizarán los trade-offs entre precisión y fairness a través de la exploración de fronteras de Pareto, validando la viabilidad del enfoque propuesto.

1. Motivación

En los últimos años, el uso de modelos de aprendizaje automático en la toma de decisiones ha generado preocupación debido a los sesgos que pueden derivarse de los datos y los algoritmos utilizados [12]. Los modelos de clasificación automática juegan un papel crucial en la toma de decisiones en ámbitos como la sanidad, el crédito financiero y la justicia penal. En particular, los clasificadores bayesianos, aunque eficientes y robustos, pueden mostrar disparidades en la clasificación cuando se aplican a conjuntos de datos con atributos sensibles como género, etnicidad o edad. Esto ha evidenciado problemas relacionados, ya que se puede derivar en decisiones discriminatorias. Por ello, la comunidad científica y la sociedad demanda el desarrollo de modelos más justos, capaces de equilibrar precisión y equidad.

Dentro de la familia de clasificadores bayesianos, el Average One-Dependence Estimators (AODE) [15] ha destacado en rendimiento frente al clásico Naïve Bayes (NB) [11] y otros representantes de la familia semi-NB que también relajan la independencia condicional entre atributos [1]. Inicialmente, en el clasificador AODE, todos los SPODEs aportan de manera equitativa. Sin embargo, podría aplicarse una técnica de ponderado donde se pudiera variar la importancia de los diferentes Super Parent One-Dependence Estimators (SPODEs) a la hora de realizar la clasificación. Esta estrategia de pesado va más alla de ponderar la importancia que cada variable tiene en el modelo final, ya que lo que se pondera es la importancia de un sub-modelo que incluye a todas las variables y modela explícitamente relaciones bivariadas entre algunas de ellas dada la clase. Ajustar estos pesos adecuadamente puede ser clave para mejorar tanto la calidad de la clasificación como cuán 'justa' esta es.

*Julia.Flores@uclm.es

Preprint. Under review.

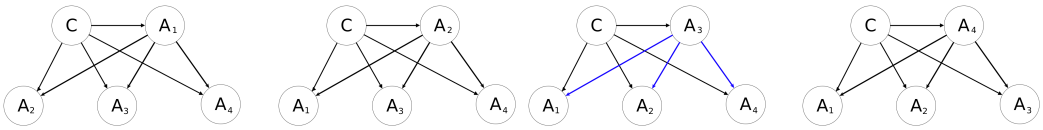

Figura 1: Figura donde se presentan los 4 SPODEs para una AODE con cuatro variables predictoras o atributos A_i.

Aunque existen variantes de AODE con pesos [9], hasta nuestro conocimiento nunca se han optimizado para propósitos de equidad. Este trabajo busca, por tanto, explorar cómo algoritmos evolutivos y optimización multiobjetivo pueden utilizarse para encontrar la mejor parametrización de pesos en AODE, maximizando métricas de rendimiento sin comprometer medidas de fairness.

2. Hipótesis de trabajo principal

En la literatura podemos encontrar distintos enfoques para abordar aspectos de equidad en el clasificador NB, como p.e. los centrados en garantizar la independencia de la clasificación respecto al valor de un atributo sensible [4], el uso de más de una variable sensible [3] o en el descubrimiento y eliminación de patrones discriminativos [6]. Sin embargo, no hemos encontrado ningún estudio centrado en el clasificador AODE [15], uno de los denominados clasificadores semi-NB más eficientes y eficaces.

AODE es un clasificador de los llamados generativos, puesto que en su aprendizaje trata de modelar la distribución de probabilidad conjunta $P(C, \mathbf{A})$ en lugar de $P(C|\mathbf{A})$. AODE trata a todas las variables por igual, sin considerar la existencia de variables sensibles. En este trabajo partimos de la asunción de que existe la posibilidad de mejorar simultáneamente la precisión y la equidad en clasificadores AODE mediante la optimización de los pesos asignados a cada SPODE, o al menos mejorar en equidad sin perjudicar en exceso la precisión. Puesto que se dispone de algoritmos evolutivos dirigidos a la optimización multiobjetivo [2], podríamos aplicarlo a nuestro tema de estudio, adaptándolo también a Naive Bayes, para compararlo con un clasificador bayesiano más sencillo. Diferentes estrategias de optimización pueden llevar a distintos trade-offs entre precisión y fairness, lo que sugiere la necesidad de analizar el frente de Pareto para seleccionar la mejor configuración según el caso de uso.

Consideramos que es un ámbito de estudio potencialmente interesante, y que es diferente al pesado de variables [14] o de instancias [13], puesto que en el AODE se combinan SPODEs (ver Figura 1), resultando en una agregación de múltiples clasificadores con restricciones estructurales para facilitar el aprendizaje y limitar su complejidad.

Para determinar la clasificación de una instancia se emplea la siguiente expresión:

$$\hat{c} = \arg\max_{C} \sum_{i=1}^{n} P(C)P(A_i \mid C) \prod_{\substack{j=1 \\ j\neq i}}^{n} P(A_j \mid A_i, C)$$

donde:

- $P(C)$ es la probabilidad a priori de la clase.
- $P(A_i \mid C)$ es la probabilidad del superpadre dado la clase, para cada uno de los SPODEs.
- $P(A_j \mid A_i, C)$ es la probabilidad condicional de los otros atributos dado el superpadre y la clase.

El enfoque que proponemos, buscará integrar un vector de pesos optimizado: $w_1, w_2, \ldots w_n$ tal que la clasificación sea:

$$\hat{c} = \arg\max_{C} \frac{1}{\sum_{k=1}^{n} w_k} \times \sum_{i=1}^{n} w_i \cdot P(C)P(A_i \mid C) \prod_{\substack{j=1 \\ j\neq i}}^{n} P(A_j \mid A_i, C)$$

donde en definitiva cada SPODE será dotado de una importancia. Pretendemos usar pesos numéricos, y en principio los w_i estarían en el intervalo [0,1].

3. Objetivos

El principal propósito es diseñar e implementar un método basado en optimización multiobjetivo con algoritmos evolutivos para ajustar los pesos en un clasificador AODE, con el fin de mejorar tanto la precisión como la equidad en la clasificación. Para ello, más específicamente, pretendemos:

1. Implementar un esquema de optimización de pesos para AODE que permita mejorar métricas de fairness sin comprometer significativamente la precisión.
2. Comparar el rendimiento del AODE ajustado con el AODE estándar y el Naïve Bayes como baseline.
3. Evaluar el impacto del ajuste de pesos en diferentes datasets con atributos sensibles.
4. Analizar los trade-offs entre precisión y fairness mediante la exploración de fronteras de Pareto. Para ello consideraremos distintas medidas de equidad de entre las disponibles en la literatura:
   - *Paridad Estadística:* La probabilidad de una predicción positiva, dada la pertenencia a un grupo, debe ser igual para todos los grupos.
   - *Impacto Dispar:* La media del cociente de predicciones positivas entre cada par de grupos debe ser 1 o mayor que un porcentaje $p\,\%$ determinado.
   - *Equidad Diferencial:* Aplicación de la equidad grupal a grupos definidos por múltiples atributos sensibles superpuestos.
   - *Equidad Individual:* La diferencia en la probabilidad de los resultados entre dos individuos no debe ser mayor que la distancia de similitud entre ellos.
   - *Equidad Causal:* Uso de modelado causal para determinar el efecto de los atributos sensibles en las predicciones.

4. Metodología

Este es el procedimiento que planeamos seguir:

1. Selección de datasets.- Se trabajará con datasets de clasificación que contengan variables sensibles, como COMPAS, Adult Income y German Credit, entre otros. [10]. Se discretizarán los atributos numéricos para su uso en clasificadores bayesianos. Implementación de los modelos
2. Implementación de modelos.- Naïve Bayes se usará como baseline, proporcionando un punto de comparación sin ajuste de pesos. AODE con ajuste de pesos será el modelo principal, donde los pesos de los SPODEs se optimizarán mediante algoritmos evolutivos.
3. Optimización de pesos.- Se utilizarán algoritmos evolutivos para encontrar configuraciones óptimas de pesos. Se empleará un enfoque de optimización multiobjetivo (e.g., NSGA-II, MOEA/D) [7] para equilibrar precisión y fairness.
4. Evaluación y validación.- Se medirán métricas de precisión como accuracy, F1-score. Se evaluará fairness mediante algunas métricas como Disparate Impact, Equalized Odds, Absolute Between-ROC Area (ABROCA) o las diferencias en tasas de falsos positivos/negativos entre grupos sensibles [8]. Se analizarán los trade-offs a través de la visualización de fronteras de Pareto. Se buscará la mejor configuración propuesta mediante técnicas como el *Compromise programming* [5]
5. Análisis de resultados .- Se compararán los resultados obtenidos con los distintos modelos y algoritmos evolutivos empleados. Se evaluará la capacidad del enfoque propuesto para mejorar la equidad sin una degradación significativa del rendimiento.

Agradecimientos. Trabajo parcialmente financiado por el Gobierno de Castilla-La Mancha, la Universidad de Castilla-La Mancha y los Fondos Europeos, UE, mediante los proyectos SBPLY/21/180225/000062 y 2022-GRIN-34437. Trabajo parcialmente financiado por MICIU/AEI/10.13039/501100011033 and ERDF, EU mediante el proyecto PID2022-139293NB-C32.

Referencias

[1] Concha Bielza and Pedro Larrañaga. Discrete bayesian network classifiers: A survey. *ACM Comput. Surv.*, 47(1), July 2014.

[2] Julian Blank and Kalyanmoy Deb. Pymoo: Multi-objective optimization in python. *Ieee access*, 8:89497–89509, 2020.

[3] Stelios Boulitsakis-Logothetis. Fairness-aware naive bayes classifier for data with multiple sensitive features, 2022.

[4] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Min. Knowl. Discov.*, 21(2):277–292, 2010.

[5] Wei Chen, Margaret M Wiecek, and Jinhuan Zhang. Quality utility: a compromise programming approach to robust design. In *International design engineering technical conferences and computers and information in engineering conference*, volume 80326, page V002T02A032. American Society of Mechanical Engineers, 1998.

[6] YooJung Choi, Golnoosh Farnadi, Behrouz Babaki, and Guy Van den Broeck. Learning fair naive bayes classifiers by discovering and eliminating discrimination patterns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10077–10084, 2020.

[7] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, 6(2):182–197, 2002.

[8] Pratyush Garg, John Villasenor, and Virginia Foggo. Fairness metrics: A comparative analysis. In *2020 IEEE international conference on big data (Big Data)*, pages 3662–3666. IEEE, 2020.

[9] Liangxiao Jiang, Harry Zhang, Zhihua Cai, and Dianhong Wang. Weighted average of one-dependence estimators. *Journal of Experimental & Theoretical Artificial Intelligence*, 24(2):219–230, 2012.

[10] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3):e1452, 2022.

[11] Kevin P Murphy et al. Naive bayes classifiers. *University of British Columbia*, 18(60):1–8, 2006.

[12] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

[13] Andrija Petrović, Mladen Nikolić, Sandro Radovanović, Boris Delibašić, and Miloš Jovanović. Fair: Fair adversarial instance re-weighting. *Neurocomputing*, 476:14–37, 2022.

[14] Ricardo Salazar, Felix Neutatz, and Ziawasch Abedjan. Automated feature engineering for algorithmic fairness. *Proceedings of the VLDB Endowment*, 14(9):1694–1702, 2021.

[15] Geoffrey I Webb, Janice R Boughton, and Zhihai Wang. Not so naive bayes: aggregating one-dependence estimators. *Machine learning*, 58:5–24, 2005.