Predictor Basado en Prototipos Difusos y Clasificación No-supervisada

JIISIC-CEIS'2015

Aníbal Vásquez, Enrique Peláez y Xavier Ochoa

anibal.vasquez @cti.espol.edu.ec, epelaez@cti.espol.edu.ec, xavier@cti.espol.edu.ec

Agenda

- Introducción
- Creación de prototipos difusos
- Arquitectura propuesta
- Modelo aplicado a un caso de estudio
- Análisis de resultados
- Conclusiones

Introducción

El clustering es una técnica que permite realizar subagrupaciones, llamadas clusters en un conjunto, permitiendo clasificar a los elementos de acuerdo a una similitud entre ellos (Kaufman & Rousseeuw, 2009)

Introducción

Existe una generalización de esta técnica hacia conjuntos difusos, para así representar a cada cluster como un conjunto difuso (Bezdek, 1981)

Introducción

Esta investigación se centra en la creación de prototipos difusos, para identificar características comunes que comparten individuos en una población.

Introducción

Para validar esta técnica, se realizó un estudio en estudiantes de la carrera de Ciencias Computacionales en una universidad, para predecir el riesgo de falla en futuros estudiantes.

Creación de prototipos difusos

Las técnicas de clustering difuso pueden ser usadas para la construcción de prototipos difusos, este es método que permite extraer elementos característicos para categorías en un conjunto (Lesot, 2005)

Creación de prototipos difusos

Este método se basa en la construcción de elementos representativos en base a una noción de tipicidad asociada a cada elemento, que se define en funciń de la semejanza y similitud (Lesot et. al, 2008)

Creación de prototipos difusos

La tipicidad se define como la agregación de la semejanza y disimilitud (Rifqi, 2000), formalmente como:

$$t_{ir} = \begin{cases} \Phi\big(R_r(x_i), D_r(x_i)\big), x_i \in C_r \\ 0, x_i \notin C_r \end{cases}$$

Creación de prototipos difusos

Donde:

$$R_r(x_i) = \frac{1}{|C_r|} \sum_{y \in C_r} \rho\,(x_i, y)$$

es la semejanza

$$D_r(x_i) = \frac{1}{|X/C_r|} \sum_{z \notin C_r} \delta\,(x_i, z)$$

es la disimilitud

Creación de prototipos difusos

Mediante la tipicidad asociada a cada elemento el prototipo difuso es construido, esto como la agregación sobre los elementos que cumplen la condición de umbra (Lesot et. al, 2008), formalmente como:

$$w_r = \psi(\{x_i / t_{ir} > \tau\})$$

Creación de prototipos difusos

Fuzzy C-Means es un algoritmo de clustering difuso, los grados de membresía asociados a cada elemento son calculados y describen la pertenencia a cada cluster (Bezdek et. al, 1984), estos grados se definen como:

$$u_{ir} = \left(\sum_{s=1}^{c} \left(\frac{d_{ir}}{d_{is}} \right)^{2/(m-1)} \right)^{-1}$$

Creación de prototipos difusos

Además, este algoritmo puede ser visto como un proceso de media ponderada de forma iterativa, considerándose como una optimización para determinar la semejanza (Lesot et. al, 2005)

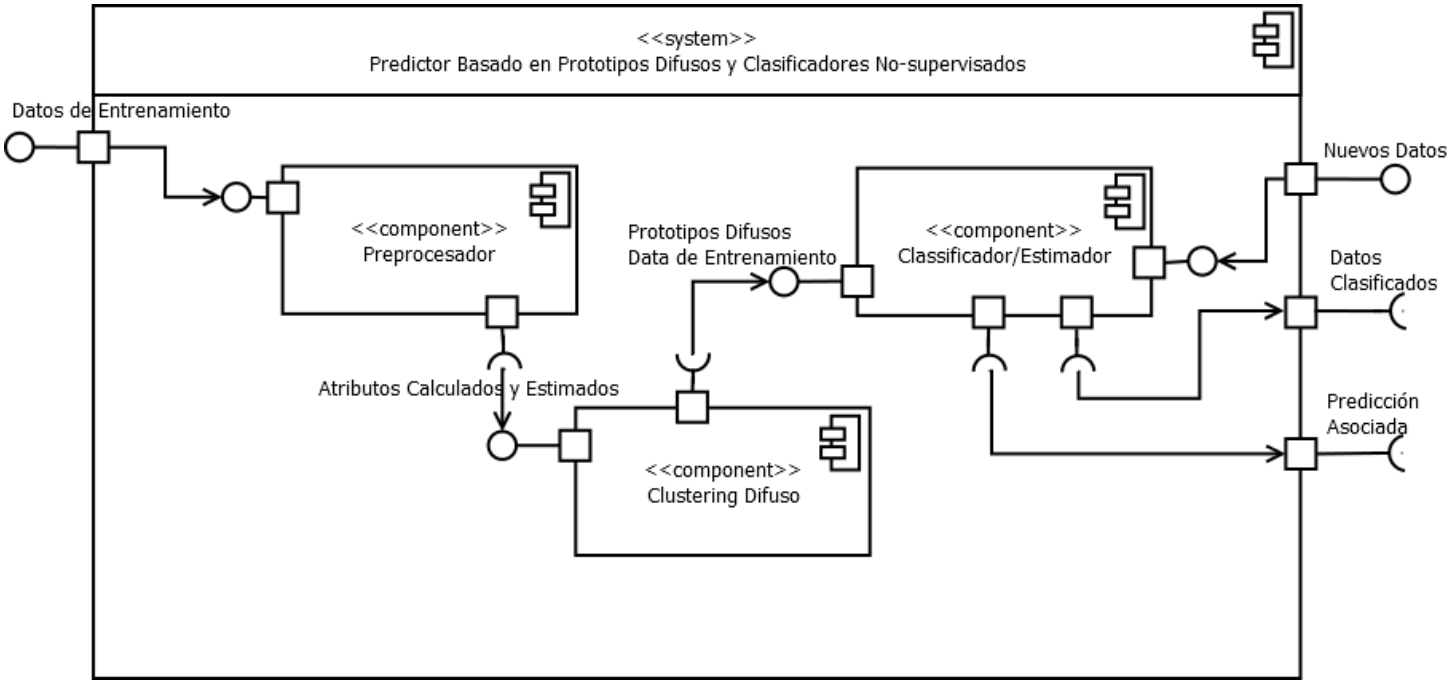
Arquitectura propuesta

El software diseñado para predicción soportada por prototipos difusos y clasificación no-supervisada, se basa en una arquitectura de componentes.

Arquitectura propuesta

Esta consta de 3 componentes principales:

- Preprocesador.
- Clustering Difuso.
- Clasificador/Estimador.



Arquitectura propuesta para predicción a través de prototipos difusos y clasificación no-supervisada.

Modelo aplicado a un caso de estudio

En años recientes la predicción del rendimiento de estudiantes también ha sido explorada a través de redes neuronales no-supervisadas, que son generalmente utilizadas para clasificar objetos en diferentes grupos (Oladokun et al., 2008) (Naser et al., 2015)

Modelo aplicado a un caso de estudio

El modelo propuesto fue aplicado a un conjunto de datos de estudiantes para predecir el riesgo de reprobación, los datos académicos de estudiantes entre 1978 y 2012 fueron usados como entrenamiento.

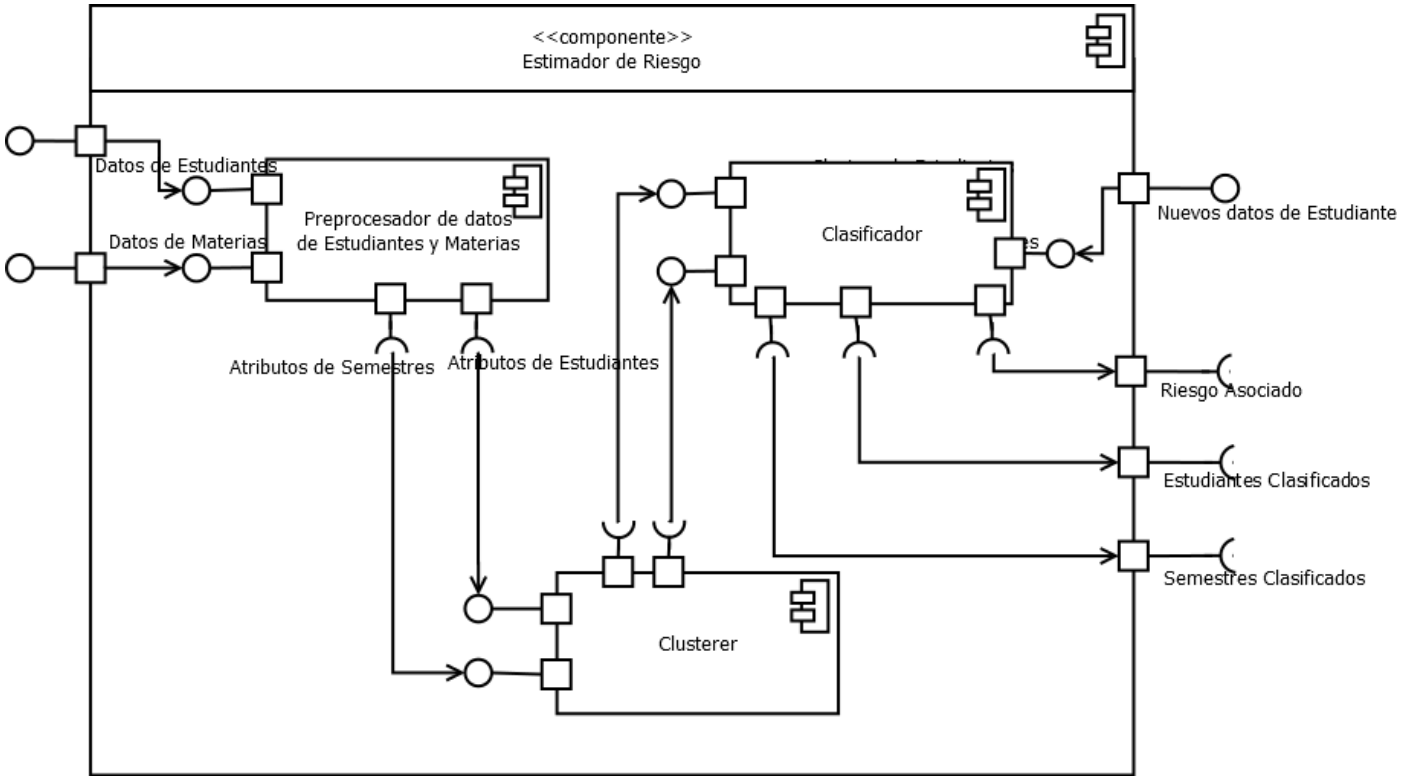


Diagrama de componentes del software de estimación de riesgo de reprobación

Modelo aplicado a un caso de estudio

El predictor de riesgo extraía de la historia académica de los estudiantes variables relacionadas con carga académica semestral (Caulkins et. al, 1996) y rendimiento de estudiantes (Mendez et. al, 2014) y asociar el riesgo de reprobación entre estudiantes similares.

Análisis de resultados

Se realizaron varios experimentos para identificar adecuadamente el número de categorías de estudiantes que serán obtenidas en el componente de clustering usando el índice de validación de Dave (Dave, 1996).

C	V_{MPC}
5	0.86077604
8	0.85675315
9	0.85385448
4	0.85272689
7	0.85163307
6	0.85047072
3	0.84808947
10	0.84545792
11	0.84428366
12	0.83817088

Número de clusters *C* sugeridos de acuerdo al índice de validación de Dave

Factor 1: Ingeniería Básica
Factor 2: Interacción con el Cliente
Factor 3: Temas Avanzados de CS
Factor 4: Programación
Factor 5: Ingeniería Eléctrica

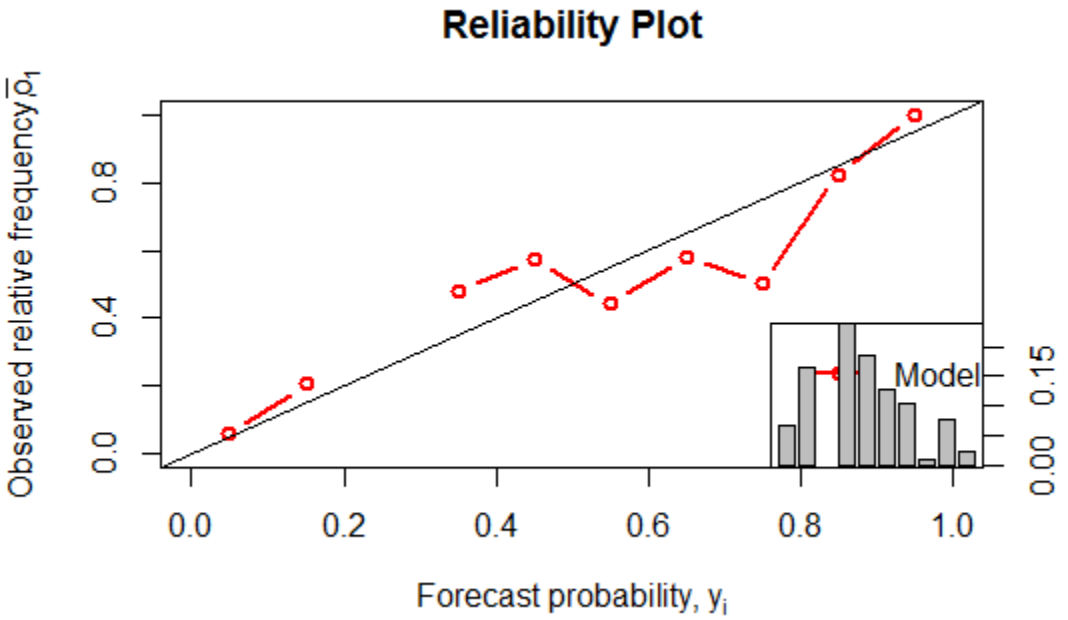
Número de clusters *C* sugeridos de acuerdo al índice de validación de Dave

Análisis de resultados

Para analizar el nivel de confiabilidad se llevó a cabo una verificación mediante el Score de Brier (Brier, 1950), este experimento es realizado para probar el nivel de certeza que puede manejar el modelo implementado.

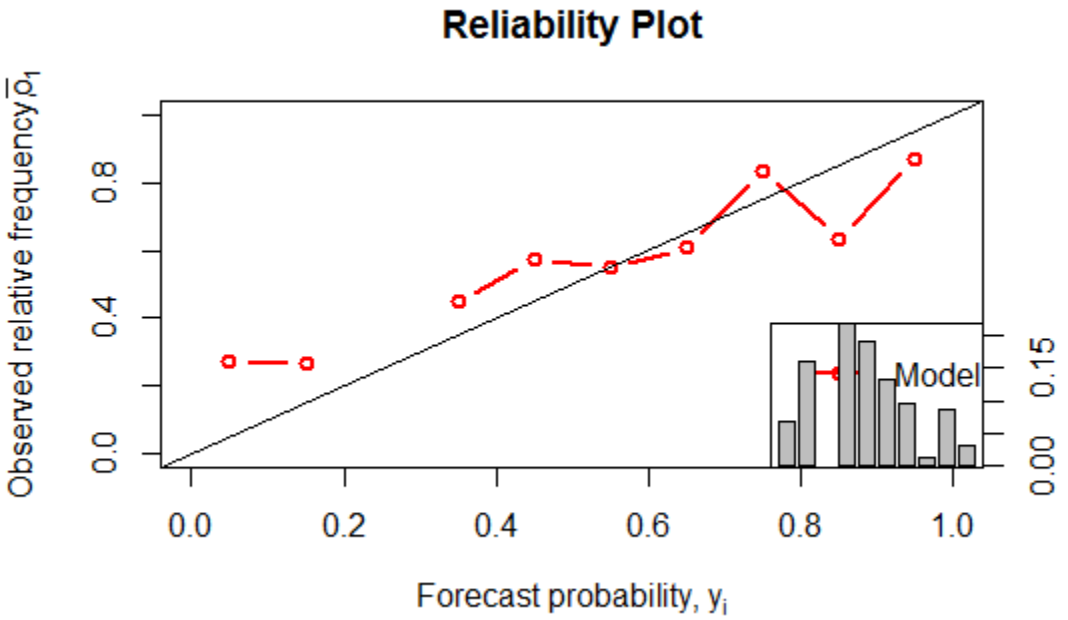
Análisis de resultados

En este experimento se entrenó el modelo con los datos de los estudiantes hasta el año 2012, la estimación se realizó para los términos I y II del año 2013 y se contrastó con la observación de reprobación real en dichos términos.



Brier Score (BS) = 0.2164
Skill Score = 0.1304
Reliability = 0.0101
Resolution = 0.04256
Uncertainty = 0.2488

Confiabilidad para la predicción de reprobación en el término 2013-I



Brier Score (BS) = 0.2422
Skill Score = 0.03113
Reliability = 0.01486
Resolution = 0.02264
Uncertainty = 0.2499

Confiabilidad para la predicción de reprobación en el término 2013-I

Conclusiones

Los algoritmos de clustering difuso permiten construir prototipos a través de la identificación de tipicidad.

Estos prototipos representan los elementos más distintivos de cada grupo de manera que pueden ser utilizados para clasificar nuevos datos.

Conclusiones

La arquitectura propuesta, aplicada a un caso de estudio, permitió predecir el riesgo de reprobación de estudiantes, con una certeza superior al 75%.

Esto brindaría soporte a estudiantes en su toma de decisiones, antes de registrarse en los cursos en un semestre.

Referencias

- (Kaufman & Rousseeuw, 2009) Kaufman, L., & Rousseeuw, P. J. Finding groups in data: an introduction to cluster analysis (Vol. 344). John Wiley & Sons (2009).
- (Bezdek, 1981) Bezdek, J. C. (1981). Pattern recognition with fuzzy objective function algorithms. Kluwer Academic Publishers (1981).
- (Lesot, 2005) Lesot, M. J. Similarity, typicality and fuzzy prototypes for numerical data. In 6th European Congress on Systems Science, Workshop Similarity and resemblance Vol. 94. (2005) 95-96.
- (Rifqi, 2000) Rifqi, M., Berger, V., & Bouchon-Meunier, B. Discrimination power of measures of comparison. Fuzzy sets and systems, 110(2), (2000) 189-196.

Referencias

- (Lesot et. al, 2008) Lesot, M. J., Rifqi, M., & Bouchon-Meunier, B. Fuzzy prototypes: From a cognitive view to a machine learning principle. InFuzzy Sets and Their Extensions: Representation, Aggregation and Models. Springer Berlin Heidelberg (2008) 431-452.
- (Bezdek et. al, 1984) Bezdek, J. C., Ehrlich, R., & Full, W. FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences, 10(2) (1984) 191-203.
- (Lesot et. al, 2005) Lesot, M. J., Mouillet, L., & Bouchon-Meunier, B. Fuzzy prototypes based on typicality degrees. In Computational Intelligence, Theory and Applications (pp. 125-138). Springer Berlin Heidelberg (2005).

Referencias

- (Oladokun et al., 2008) Oladokun, V. O., Adebanjo, A. T., & Charles-Owaba, O. E. Predicting students' academic performance using artificial neural network: A case study of an engineering course. The Pacific Journal of Science and Technology (2008) 72-79.
- (Naser et al., 2015) Naser, S. A., Zaqout, I., Ghosh, M. A., Atallah, R., & Alajrami, E. Predicting Student Performance Using Artificial Neural Network: in the Faculty of Engineering and Information Technology (2015).
- (Caulkins et. al, 1996) Caulkins, J. P., Larkey, P. D., & Wei, J. Adjusting GPA to reflect course difficulty (1996).
- (Méndez et. al, 2014) Méndez, G., Ochoa, X., & Chiluiza, K. Techniques for data-driven curriculum analysis. In Proceedings of the Fourth International Conference on Learning Analytics And Knowledge. ACM (2014) 148-157.

Referencias

- (Dave, 1996) Dave, R. N. Validating fuzzy partitions obtained through c-shells clustering. Pattern Recognition Letters, 17(6), (1996) 613-623.
- (Brier, 1950) Brier, G. W. Verification of forecasts expressed in terms of probability.Monthly weather review, 78(1), (1950) 1-3.