Revista Cubana de Ciencias Informáticas
Vol. 17, No. 1, Mes Enero-Marzo, 2023
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 45-59

Tipo de artículo: Artículo originales
Temática: Reconocimiento de patrones
Recibido: 02/11/2022 | Aceptado: 24/12/2022 | Publicado: 06/02/2023

Proactive Forest: Análisis del impacto de la generalización del parámetro de diversidad

Proactive Forest: Analysis of the impact of the generalization of the diversity parameter

Nayma Cepero Pérez 0000-0003-3808-8135[1*]
Mailyn Moreno Espino 0000-0002-7613-3382[2]
Milton García Borroto 0000-0002-3154-177X[3]
Eduardo F. Morales 0000-0002-7618-8762[4]

[1,2,3]Universidad Tecnológica de la Habana José Antonio Echeverría,CUJAE, calle 114 No. 11901, e/ Ciclovía y Rotonda, Marianao, Cuba.
[4]Instituto Nacional de Astrofísica Óptica y Electrónica, INAOE, Luis Enrique Erro 1,Sta María Tonanzintla, 72840 San Andrés Cholula, Puebla, México, emorales@inaoep.mx

[*]Autor para correspondencia: (ncepero@ceis.cujae.edu.cu)

RESUMEN

La facilidad para interpretar las predicciones realizadas por un modelo aprendido constituye una de las ventajas que hacen de los árboles de decisión, una de las técnicas más efectivas a la hora de enfrentar una tarea de minería de datos. Las predicciones realizadas por muchos árboles de decisión pueden ser combinadas con el objetivo de mejorar la decisión final, de esta idea surge el concepto de bosques de decisión. Es condición necesaria para construir un bosque de decisión, que los árboles individuales tengan un alto poder predictivo y al mismo tiempo sean diferentes entre ellos. Esta diferencia es conocida como diversidad del bosque de decisión, conseguirla no es un proceso trivial. Los algoritmos de bosques de decisión más empleados utilizan aleatoriedad en el proceso de construcción de cada árbol para obtener diversidad; sin embargo, el uso de la aleatoriedad no siempre garantiza obtener una diversidad adecuada. Proactive Forest es un algoritmo constructor de bosques de decisión que introduce un mecanismo de control de aleatoriedad a partir de la definición

Revista Cubana de Ciencias Informáticas
Vol. 17, No. 1, Mes Enero-Marzo, 2023
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 45-59

de una función de actualización de las probabilidades con las que se utilizan los atributos, uno de los elementos más importantes es el parámetro de diversidad que se definió como 0.1 inicialmente. El objetivo de este trabajo es analizar el uso de un único valor del parámetro de diversidad para todas las bases de datos. En los resultados se demuestra que no es correcto generalizar un valor de diversidad, ya que la eficacia se afecta según el valor que se use.

Palabras clave: eficacia; diversidad; bosques de decisión; Proactive Forest; parámetro de diversidad.

ABSTRACT

The ease of interpreting the predictions made by a learned model is one of the advantages that make decision trees one of the most effective techniques when facing a data mining task. The predictions made by many decision trees can be combined in order to improve the final decision, from this idea arises the concept of decision forests. It is a necessary condition for building a decision forest that the individual trees have a high predictive power and at the same time are different from each other. This difference is known as decision forest diversity, and achieving it is not a trivial process. The most commonly used decision forest algorithms use randomization in the process of constructing each tree to obtain diversity; however, the use of randomization does not always guarantee obtaining adequate diversity. Proactive Forest is a decision forest construction algorithm that introduces a randomness control mechanism based on the definition of an update function of the probabilities with which the attributes are used, one of the most important elements is the diversity parameter that was initially defined as 0.1. The objective of this work is to analyze the use of a single value of the diversity parameter for all the databases. The results show that it is not correct to generalize a diversity value, since the effectiveness is affected depending on the value used.

Keywords: accuracy; diversity ; decision forest ; Proactive Forest ; diversity parameter.

Introducción

El aprendizaje *ensemble* se refiere a la generación y combinación de múltiples modelos para resolver una tarea de aprendizaje en particular. Esta metodología imita la naturaleza de los seres humanos de buscar varias

Revista Cubana de Ciencias Informáticas
Vol. 17, No. 1, Mes Enero-Marzo, 2023
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 45-59

opiniones antes de tomar una decisión crucial, como escoger un determinado tratamiento médico por ejemplo (Lundberg et al., 2020). El principio fundamental del *ensemble* es darle un peso a cada una de las opiniones individuales y combinarlas todas con el objetivo de obtener una decisión que sea mejor que las obtenidas por cada uno de ellas por separado (Polikar, 2006).

Existen diferentes métodos de construcción de bosques de decisión usando un método ensemble general como Adaboost (Freund et al., 1996), el cual puede ser utilizado con cualquier algoritmo de aprendizaje base o usando un método *ensemble* específicamente diseñado para la construcción de bosques de decisión como Random Forest (Breiman, 2001). Los algoritmos más utilizados en la constucción de bosques de decisión son:

- Bagging: escoge aleatoriamente las instancias que forman parte del conjunto de entrenamiento con el cual se construye cada árbol (Breiman, 1996).
- Random Subspaces: construye cada árbol utilizando una muestra aleatoria de los atributos (Ho, 1998).
- Random Forest: tiene dos procesos que inyectan aleatoriedad, el primero es escoger aleatoriamente las instancias que forman parte del conjunto de entrenamiento con el cual se construye cada árbol y el segundo consiste en escoger el mejor *split* posible entre una muestra de los atributos en vez de considerar todos los atributos (Breiman, 2001).

Se ha mostrado repetidamente que el aprendizaje *ensemble* mejora el poder predictivo de un modelo individual, este trabaja particularmente bien cuando son usados árboles de decisión como modelos bases (Polikar, 2006). De esta idea, de combinar el aprendizaje *ensemble* con los árboles de decisión surge la noción de los bosques de decisión (Lundberg and Lee, 2017) .

En la construcción de un buen bosque de decisión se deben cumplir los principios de diversidad y poder predictivo. Un bosque es diverso cuando cada árbol de decisión individual es lo suficientemente diferente de los demás (Mitchell, 1980; Molnar, 2018). A su vez, el poder predictivo de cada árbol debe ser el mejor posible o al menos mejor que un modelo aleatorio. La combinación de ambos principios tienen como objetivo hacer los bosques de decisión más eficaces.

Existen diferentes acercamientos para lograr diversidad en un *ensemble* (Ali and Pazzani, 1995; Lundberg et al., 2018) por ejemplo: manipular el conjunto de entrenamiento. En este caso se varía la entrada que es usada por el algoritmo; cada árbol es entrenado con un conjunto diferente de entrenamiento. Este método es útil para modelos en los cuales pequeños cambios en el conjunto de entrenamiento pueden provocar grandes cambios en ellos.

Existen dos maneras obvias de dividir el conjunto de datos original, horizontal y verticalmente. En la división horizontal el conjunto de datos original es dividido en muchos subconjuntos que contienen los mismos atributos que el original, cada uno conteniendo un subconjunto de las instancias del original (Chawla et al., 2004). En la partición vertical el conjunto original es dividido en muchos subconjuntos que tienen la misma cantidad de instancias que el conjunto original, cada uno conteniendo un subconjunto de los atributos del original (Rokach, 2008).

Otra manera de lograr diversidad es manipulando el algoritmo de aprendizaje. En este acercamiento, se manipula la manera en la que el algoritmo de construcción de árboles es utilizado. Una variante es alternar la forma en la que el algoritmo atraviesa el espacio de hipótesis, logrando así, que se construyan árboles de decisión diferentes (Brown et al., 2005). También se puede buscar diversidad mediante la manipulación de los parámetros del algoritmo. Estos algoritmos normalmente son controlados por un conjunto de parámetros, tales como la altura del árbol y la cantidad mínima de instancias para una hoja. Ejecutar el algoritmo con diferentes configuraciones de estos parámetros produce árboles diferentes (Lin and Chen, 2012).

Luego de analizar el comportamiento de los algoritmos existentes para construir bosques de decisión, se identificó que en muchas ocasiones los atributos de mayor relevancia aparecen en la mayoría de los árboles construidos, lo que hace que los árboles construidos no sean tan diferentes y por tanto, la diversidad del bosque de decisión se ve afectada (Fan, 2022). Adicionalmente se inyectan uno o varios procesos aleatorios en la construcción del bosque, lo que trae como consecuencia que no se puedan entender del todo los resultados obtenidos y que se afecte la eficacia de la clasificación.

Con el objetivo de controlar la diversidad que se genera en el bosque de decisión se identificaron tres aspectos del proceso de construcción de los árboles en los cuales se podría introducir un componente de control: intancias, *split* y atributos. El trabajo con las intancias fue explorado en (Freund et al., 1996), con los *split* y los atributos se trabaja en (García-Borroto et al., 2015). En el trabajo presentado en (García-Borroto et al., 2015) se introduce un método de control de diversidad en el cual, los atributos que hayan formado parte de un árbol de decisión no pueden volver a utilizarse para construir los árboles de decisión siguientes. Este algoritmo tiene un inconveniente, para bases de datos con pocos atributos, el conjunto de árboles que puede construirse es muy pequeño. Además, ha demostrado ser menos eficaz que los métodos para construir bosques de decisión que generan diversidad a través de los *split*. Por ello, se consideró que podría realizarse una mayor investigación en la cual los atributos fueran el centro de atención.

Teniendo en cuenta lo planteado en estos trabajos se desarrolló el algoritmo Proactive Forest presentado en (Cepero-Pérez et al., 2018). El algoritmo Proactive Forest construye el bosque de manera secuencial. En cada iteración se construye un árbol de decisión, luego se calcula la importancia de los atributos en el bosque

Revista Cubana de Ciencias Informáticas
Vol. 17, No. 1, Mes Enero-Marzo, 2023
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 45-59

construido hasta el momento y se actualizan las probabilidades de los atributos de acuerdo a su importancia. De esta forma se logra balancear el impacto de los atributos más relevantes en el bosque de decisión y con ello se genera una diversidad más controlada (Cepero-Pérez et al., 2018).
Se realizaron varios experimentos en los cuales se determinaron los factores que estarían presentes en la función de actualización y por tanto afectarían cómo se realiza la actualización de probabilidades. Los factores identificados son (Cepero-Pérez et al., 2018):

- alpha, es el parámetro de diversidad que define cuán rápido se quiere que las probabilidades de los atributos más importantes disminuyan.
- iter/cant árboles, es el control de construcción que se utliza para que la actualización de las probabilidades se haga en correspondencia con el proceso de construcción del bosque.

Inicialmente el valor de alpha fue definido en 0.1, pero del mismo modo en que no se puede definir que un clasificador sea mejor que otro porque en depenencia del tipo de problema uno obtiene mejores resultados que otro, se plantea la necesidad de analizar experiementalmente el impacto de alpha en la eficacia de la clasificación. Partiendo de la idea de que no es correcto fijar este valor para todos los problemas.

Análisis sobre la generalización del parámetro de diverisdad

Para realizar los experimentos se utilizaron bases de datos de diversas características extraídas del repositorio UCI Machine Learning, las cuales son consideradas estándares para evaluar soluciones de aprendizaje automático (Dheeru and Karra Taniskidou, 2017). Las bases de datos contienen desde 3 hasta 64 atributos, desde 2 hasta 26 clases y desde 106 hasta 20000 instancias. En la Tabla 1 se presentan con más detalles las características de las bases de datos utlizadas.
En los análisis se utiliza la validación cruzada para evaluar la eficacia de la clasificación, en el impacto de la generalización del paramétro de diversidad del algoritmo Proactive Forest, con k-iteraciones (k = 10), en la que los datos se separan en k subconjuntos, uno de los cuales se utiliza como conjunto de prueba y los restantes k-1 como datos de entrenamiento con el objetivo de que el proceso de validación sea más robusto. Esto garantiza que los resultados son independientes de la partición entre los datos de entrenamiento y los datos de prueba.

Revista Cubana de Ciencias Informáticas
Vol. 17, No. 1, Mes Enero-Marzo, 2023
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 45-59

Tabla 1 - Descripción de las bases de datos utilizadas.

Base de Datos	Instancias	Atributos	Clases	Bases de Datos	Instancias	Atributos	Clases
Balance scale	625	4	3	Molecular	106	58	2
Car	1728	6	4	Nursery	12960	8	5
Cmc	1473	9	3	Optdigits	5620	64	10
Credit-g	1000	20	2	Page blocks	5473	10	5
Diabetes	768	8	2	Pendigits	10992	16	10
Ecoli	336	7	8	Segment	2310	19	7
Flags religion	194	29	8	Solar flare 1	323	12	6
Glass	214	9	6	Solar flare 2	1066	12	6
Haberman	306	3	2	Sonar	208	60	2
Heart-statlog	270	13	2	Spambase	4601	57	2
Ionosphere	351	34	2	Tae	151	5	3
Iris	150	4	3	Vehicle	846	18	4
KR-vs-KP	3196	36	2	Vowel	990	13	11
Letter	20000	16	26	Wdbc	569	30	2
Liver	345	6	2	Wine	178	13	3
Lymph	148	18	4	-	-	-	-

Experimento: Análisis general del comportamiento de la eficacia

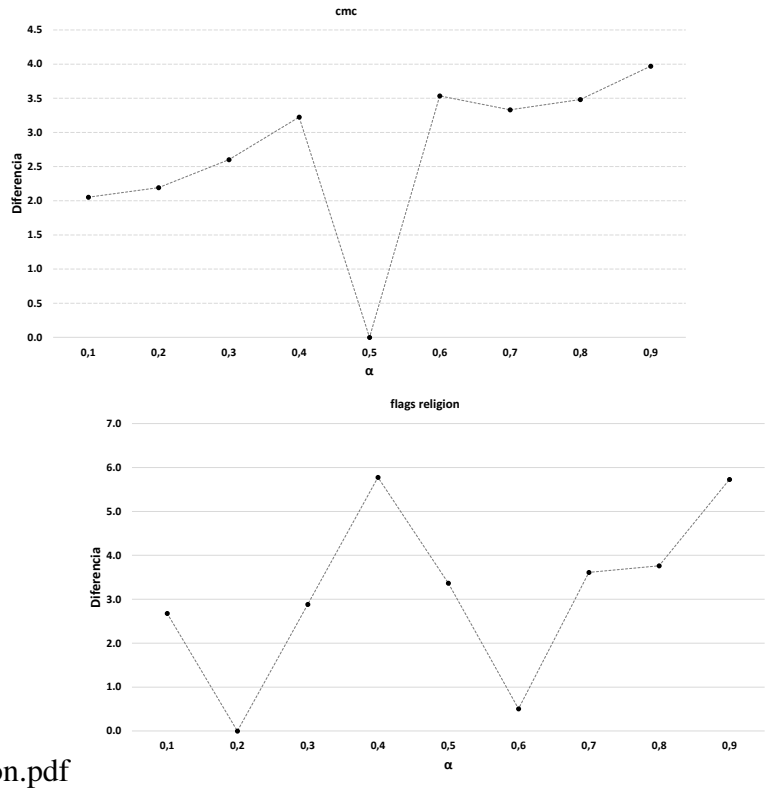
El objetivo de este experimento es demostrar la variabilidad de la eficacia generada en dependencia del parámetro de alpha que se utilice. Para la ejecución del experimento se utilizaron las bases de datos descritas anteriormente, para cada una se calculó la eficacia de su clasificación utilizando los valores de alpha desde 0.1 hasta 0.9. En el siguiente pseudocódigo se muestra la secuencia de pasos a seguir para la ejecución del experiemento:

entrada bases de datos a clasificar(**datos**)
salida modelo de clasificación(**modelo**)
alpha

1. mientras (**alpha** $\leq$ 0.9) **alpha** + 0.1
2. construir Proactive Forest (**alpha**,**datos**)
3. guardar **eficacia**
4. fin

Revista Cubana de Ciencias Informáticas
Vol. 17, No. 1, Mes Enero-Marzo, 2023
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 45-59

Posteriormente se seleccionó el mejor valor de eficacia generado por cada valor en cada base de datos y se calcularon las diferencias entre ese valor y los demás. En las Figuras **??** se presentan una muestra de estas bases de datos, que evidencian la diferencia entre usar un valor de alpha u otro.



religion.pdf

Fig. 1 - cmc y flags religion

Resultados y discusión

El estudio experimental desarrollado permitió comprender el impacto real que tiene la definición del parámetro de diversidad en la eficacia de la clasificación. A continuación se presentan los pricipales resultados obtenidos:

1. El comportamiento óptimo de alpha está estrechamente relacionado con las características de los datos, en función del conjunto de datos es mejor definir un valor u otro para alpha, cada problema de clasificación es diferente y la forma de resolverlo por tanto lo es; de ahí la importancia de definir un valor correcto de alpha en función del problema de clasificación a resolver.

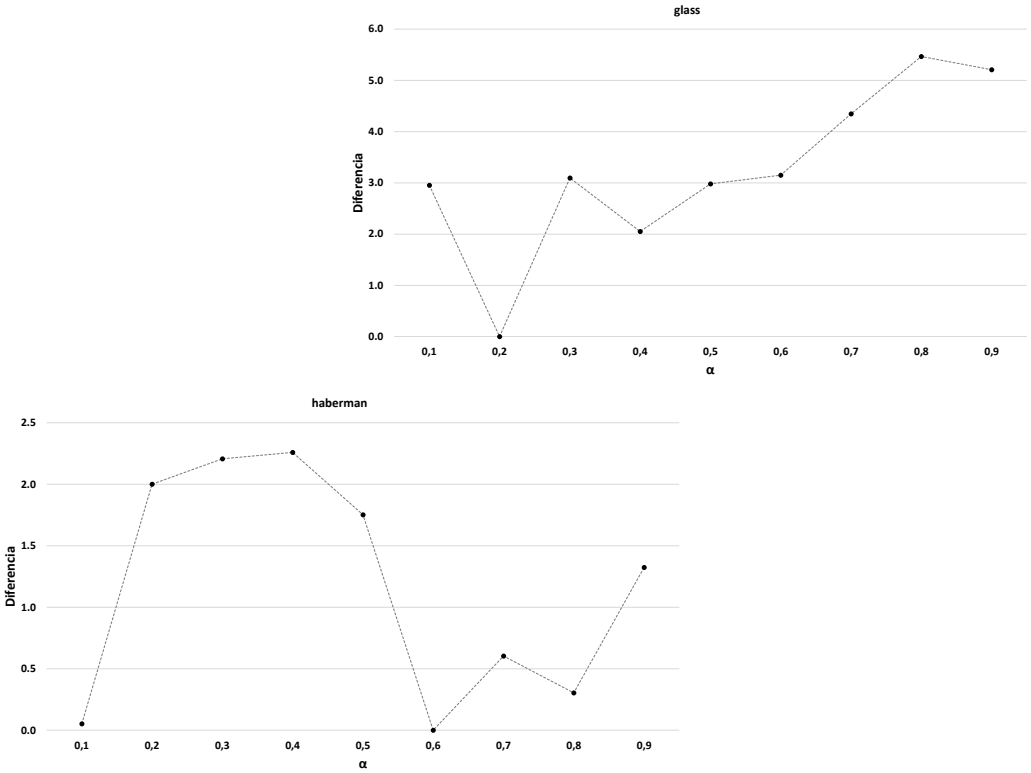


Fig. 2 - glass y haberman

Revista Cubana de Ciencias Informáticas
Vol. 17, No. 1, Mes Enero-Marzo, 2023
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 45-59

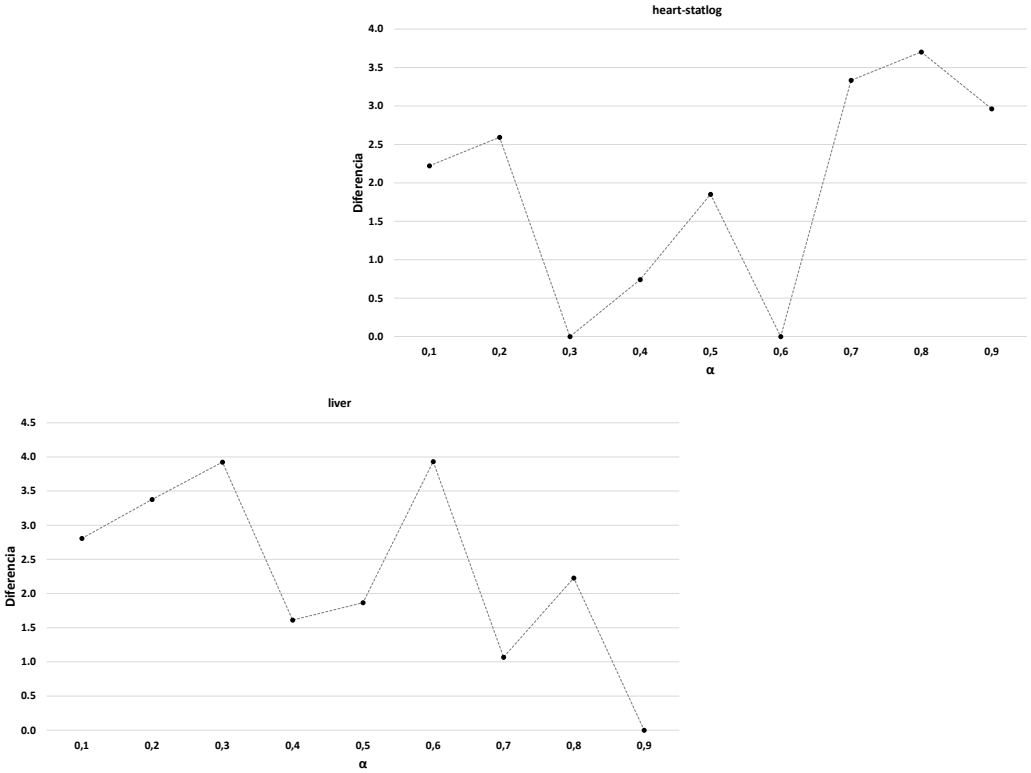


Fig. 3 - heart-statlog y liver

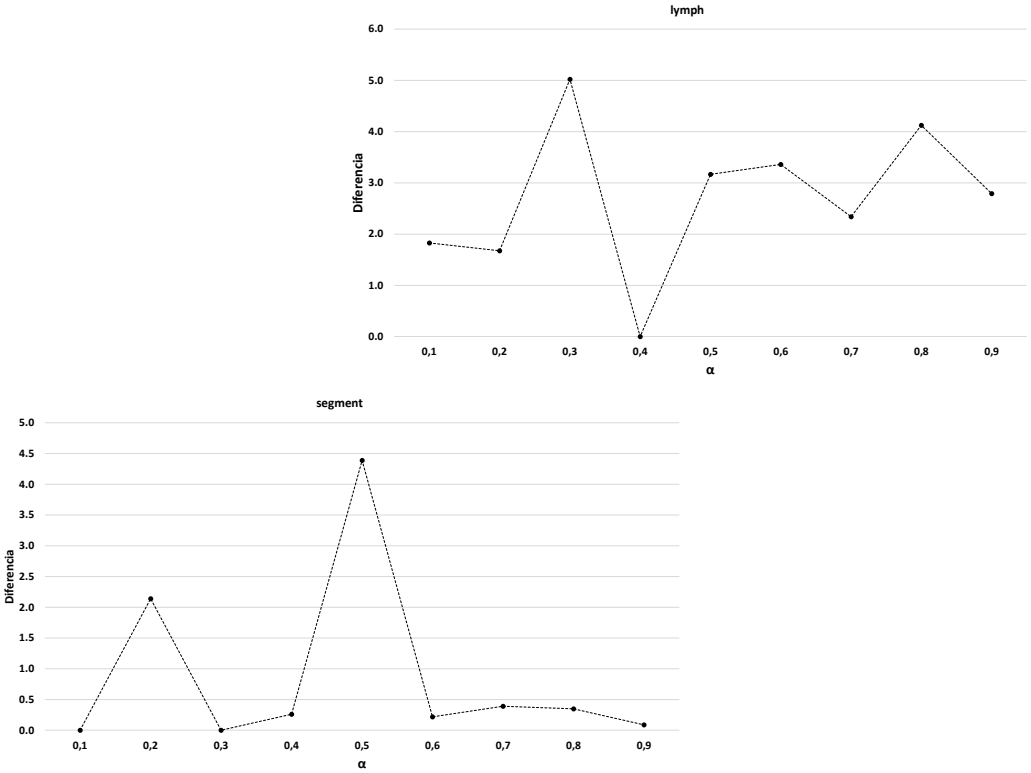
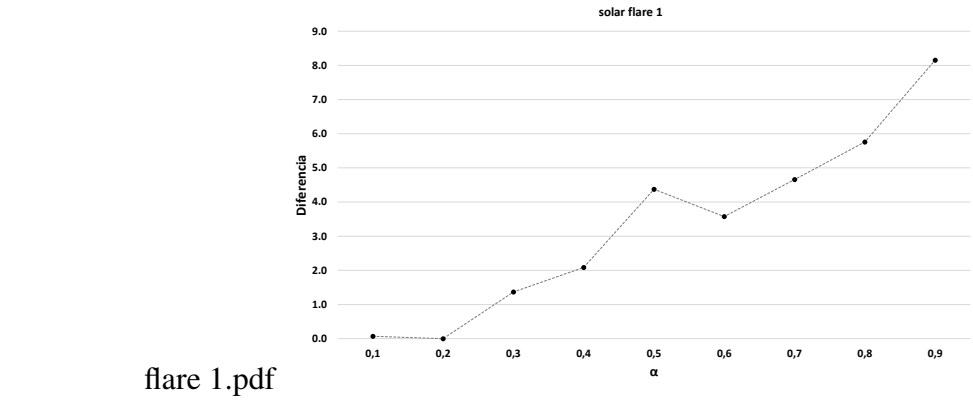
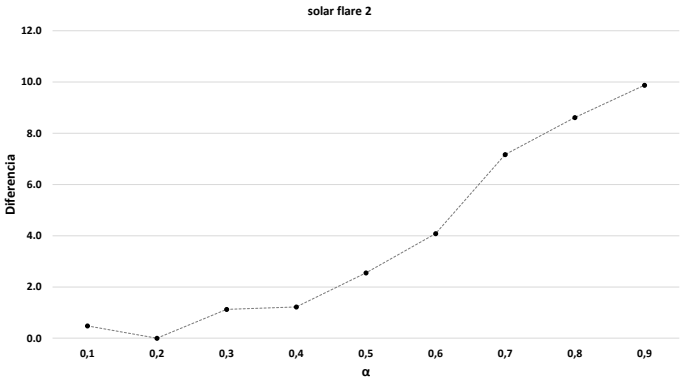


Fig. 4 - lymph y segment

Revista Cubana de Ciencias Informáticas
Vol. 17, No. 1, Mes Enero-Marzo, 2023
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 45-59



flare 1.pdf



flare 2.pdf

Fig. 5 - solar flare 1 y solar flare 2

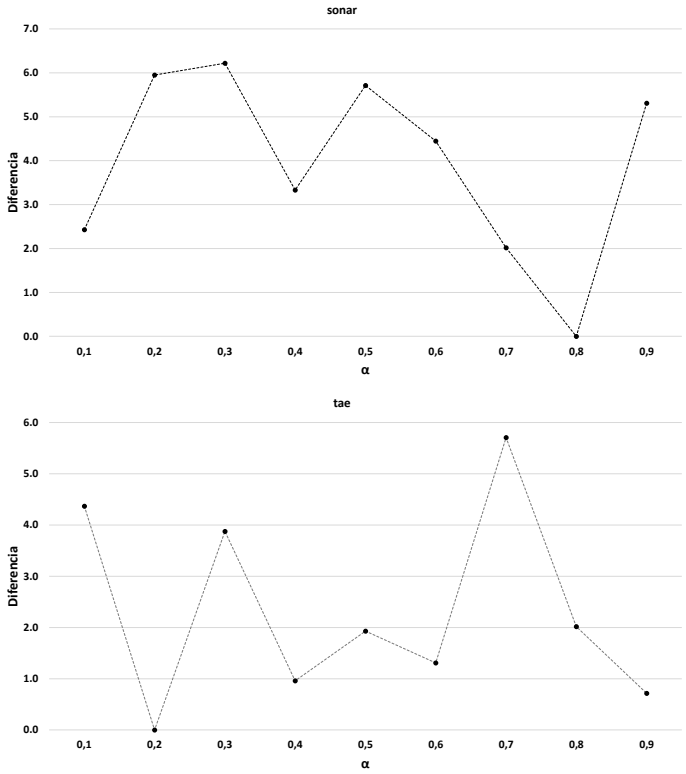


Fig. 6 - sonar y tae

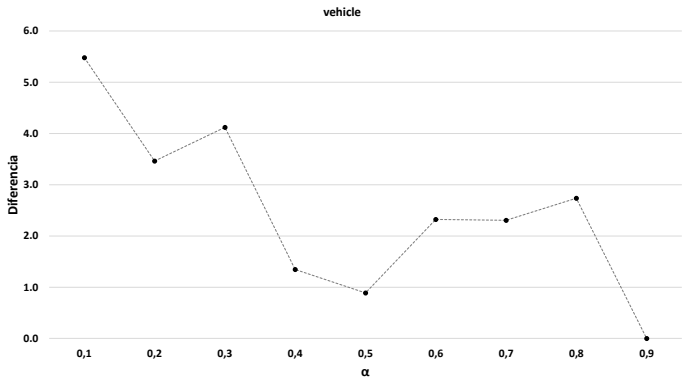


Fig. 7 - vehicle

2. Se demostró que cada valor de alpha analizado obtiene en al menos una base de datos el valor máximo de la eficacia.

3. Definir un valor general para alpha como parámetro de diversidad no es correcto.

Conclusiones

Para construir bosques de decisión eficaces se debe lograr que cada árbol de decisión individual tenga el mejor poder predictivo posible y que a su vez sea lo suficientemente diferente de los demás. La generación de la diversidad y la eficacia de la clasificación están estrechamente relacionadas por lo que es importante lograr un balance entre ambas para obtener mejores bosques. Los estudios presentados en este trabajo permitieron determinar que la generalización de un único valor para el parámetro de diversidad del algoritmo Proactive Forest no es correcto, debido a qué cada valor tiene un comportamiento óptimo en función del conjunto de datos que está procesando. Partiendo de estos estudios se desarrollará un mecanismo que permita aprender las características de las bases de datos y en función de esas características definir de manera automática qué valor utilizar en la clasificación usando Proactive Forest.

Referencias

Kamal M Ali and Michael J Pazzani. On the link between error correlation and error reduction in decision tree ensembles. 1995.

Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996. doi: 10.1007/bf00058655.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Gavin Brown, Jeremy L Wyatt, Peter Tino, and Yoshua Bengio. Managing diversity in regression ensembles. *Journal of machine learning research*, 6(9), 2005.

Nayma Cepero-Pérez, Luis Alberto Denis-Miranda, Rafael Hernández-Palacio, Mailyn Moreno-Espino, and Milton García-Borroto. Proactive forest for supervised classification. In *International Workshop on Artificial Intelligence and Pattern Recognition*, pages 255–262. Springer, 2018.

Nitesh V Chawla, Lawrence O Hall, Kevin W Bowyer, and W Philip Kegelmeyer. Learning ensembles from bites: A scalable and accurate approach. *The Journal of Machine Learning Research*, 5:421–451, 2004.

Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Ping Fan. Random forest algorithm based on speech for early identification of parkinson's disease. *Computational Intelligence and Neuroscience*, 2022, 2022.

Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer, 1996.

Milton García-Borroto, José Fco Martínez-Trinidad, and Jesús Ariel Carrasco-Ochoa. Finding the best diversity generation procedures for mining contrast patterns. *Expert Systems with Applications*, 42(11):4859–4866, 2015. doi: https://doi.org/10.1016/j.eswa.2015.02.028.

Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):832–844, 1998. doi: https://doi.org/10.1109/34.709601.

Shih-Wei Lin and Shih-Chieh Chen. Parameter determination and feature selection for c4. 5 algorithm using scatter search approach. *Soft Computing*, 16(1):63–75, 2012. doi: 10.1007/s00500-011-0734-z.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017. URL https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, 2020. doi: 10.1038/s42256-019-0138-9.

Tom M Mitchell. *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research . . . , 1980.

Christoph Molnar. A guide for making black box models explainable. *URL: https://christophm. github. io/interpretable-ml-book*, 2018.

Robi Polikar. Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, 6(3):21–45, 2006. doi: https://doi.org/10.1109/MCAS.2006.1688199.

Lior Rokach. Genetic algorithm-based feature set partitioning for classification problems. *Pattern Recognition*, 41(5):1676–1700, 2008. doi: https://doi.org/10.1016/j.patcog.2007.10.013.

Conflicto de interés

El autor autoriza la distribución y uso de su artículo.

Contribuciones de los autores

1. Conceptualización: Nayma Cepero Pérez, Mailyn Moreno Espino, Milton García Borroto, Eduardo F. Morales.
2. Curación de datos: Nayma Cepero Pérez, Eduardo F. Morales
3. Análisis formal: Nayma Cepero Pérez, Mailyn Moreno Espino
4. Investigación: Nayma Cepero Pérez, Mailyn Moreno Espino, Milton García Borroto, Eduardo F. Morales
5. Metodología: Nayma Cepero Pérez, Mailyn Moreno Espino, Milton García Borroto, Eduardo F. Morales
6. Software: Nayma Cepero Pérez.
7. Supervisión: Mailyn Moreno Espino, Milton García Borroto, Eduardo F. Morales
8. Validación: Nayma Cepero Pérez
9. Visualización:Nayma Cepero Pérez
10. Redacción - borrador original: Nayma Cepero Pérez, Mailyn Moreno Espino, Milton García Borroto, Eduardo F. Morales
11. Redacción - revisión y edición: Nayma Cepero Pérez, Mailyn Moreno Espino, Milton García Borroto, Eduardo F. Morales