

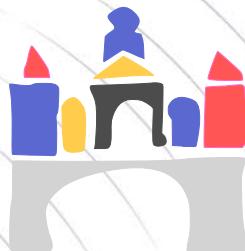
Aprendizaje supervisado. Algoritmos de construcción de ensembles y conjuntos desequilibrados

19/04/2017

José Francisco Díez Pastor

jfdpastor@ubu.es

Dept. Ingeniería civil
Universidad de Burgos



Contenidos



Topics

Introducción a la minería de datos

Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

Algoritmos de
construcción de
ensembles y conjuntos
desequilibrados

José Francisco Díez
Pastor

Topics

Introducción a la
minería de datos

Ensembles para
problemas
desequilibrados

Contribuciones al
aprendizaje de
problemas
desequilibrados

Aplicaciones

Resumen y líneas
futuras



Esta charla se centra en tres temas:

- ▶ Introducción a la minería de datos, el aprendizaje supervisado y los ensembles.
- ▶ Ensembles para conjuntos desequilibrados.
- ▶ Uso de ensembles en problemas reales.

2

Topics

Introducción a la minería de datos

Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez
Pastor

Topics

3 Introducción a la minería de datos

Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

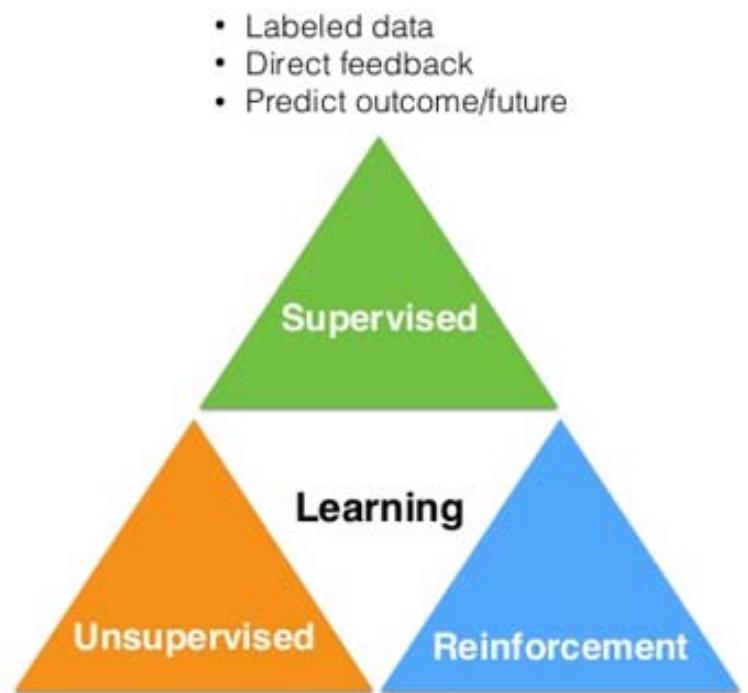
Resumen y líneas futuras

¿Qué es la minería de datos?

- ▶ La inteligencia artificial (IA) es el área de conocimiento dedicada a crear sistemas informáticos con un comportamiento inteligente.
- ▶ Dentro de la IA, la minería de datos estudia la creación de sistemas capaces de aprender por si mismos. Los algoritmos de minería de datos o aprendizaje computacional pueden dividirse en:
 - ▶ Aprendizaje no supervisado.
 - ▶ Aprendizaje con refuerzo.
 - ▶ Aprendizaje supervisado.



¿Qué es la minería de datos?



<https://adeshpande3.github.io/Deep-Learning-Research-Review-Week-2-Reinforcement-Learning>

Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

4 Introducción a la minería de datos

Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

5 Introducción a la minería de datos

Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

Aprendizaje no supervisado

- ▶ Al sistema solo se le presentan las observaciones y busca la estructura o las relaciones entre los distintos ejemplo. El tipo más popular es el clustering.

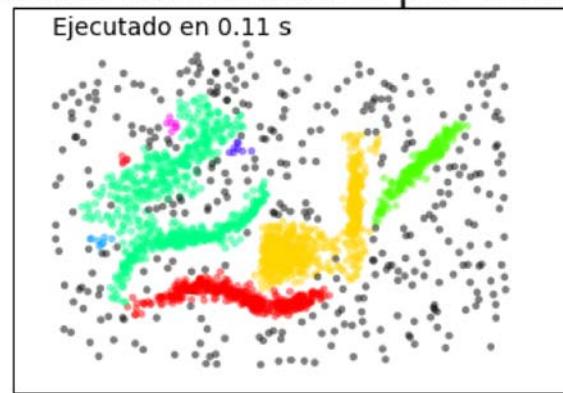
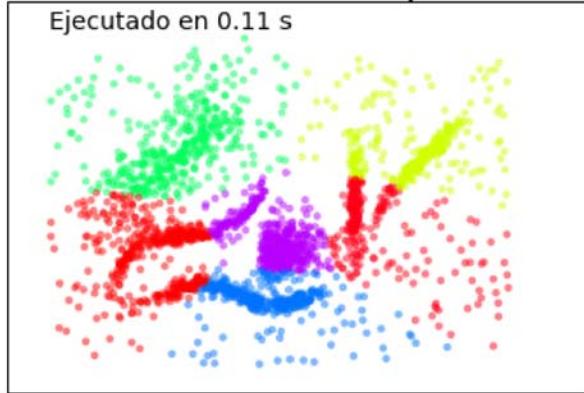
Aprendizaje con refuerzo

- ▶ En este caso al sistema se le llama agente. El agente mejora su funcionamiento a partir de interacciones con el entorno con las que recibe refuerzos positivos o negativos. Ejemplo Q-Learning



¿Qué es la minería de datos?

Clusters encontrados por KMeans Clusters encontrados por DBSCAN



- ▶ KMeans: Asigna todos los ejemplos a un cluster. Clusters esféricos. Se necesita especificar el número de clusters.
- ▶ DBSCAN: No asume que los clusteres son esfericos y no asigna todas las intancias a clusters. No es necesario especificar su número.

Algoritmos de construcción de ensambles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

6 Introducción a la minería de datos

Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

Minería de datos



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

7 Introducción a la minería de datos

Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras



Breakout and Space Invaders, 2 of the 49 Atari games used in the paper



Algoritmos de
construcción de
ensembles y conjuntos
desequilibrados

José Francisco Díez
Pastor

Topics

8 Introducción a la
minería de datos

Ensembles para
problemas
desequilibrados

Contribuciones al
aprendizaje de
problemas
desequilibrados

Aplicaciones

Resumen y líneas
futuras

Aprendizaje supervisado

- ▶ En aprendizaje supervisado se le presenta al sistema la observación junto con su clase.
- ▶ La clase puede tener muchos significados: el tipo al que pertenece el ejemplo, la mejor acción que se debe tomar en ese estado etc.
- ▶ La clase es proporcionada por un experto y el sistema tiene que aprender a relacionar los atributos de las observaciones con su clase.



Algoritmos de
construcción de
ensembles y conjuntos
desequilibrados

José Francisco Díez
Pastor

Topics

9 Introducción a la
minería de datos

Ensembles para
problemas
desequilibrados

Contribuciones al
aprendizaje de
problemas
desequilibrados

Aplicaciones

Resumen y líneas
futuras

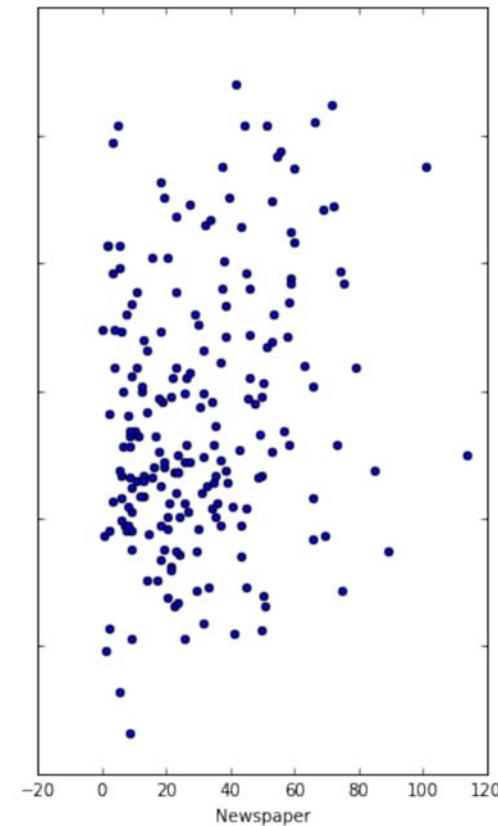
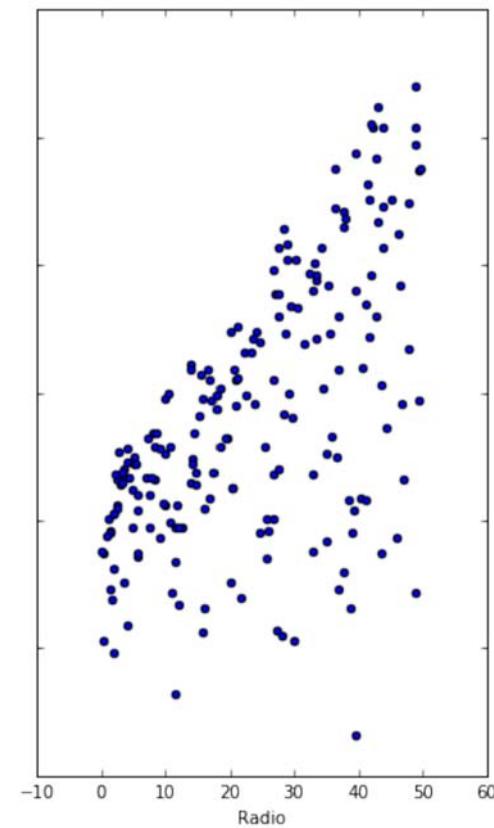
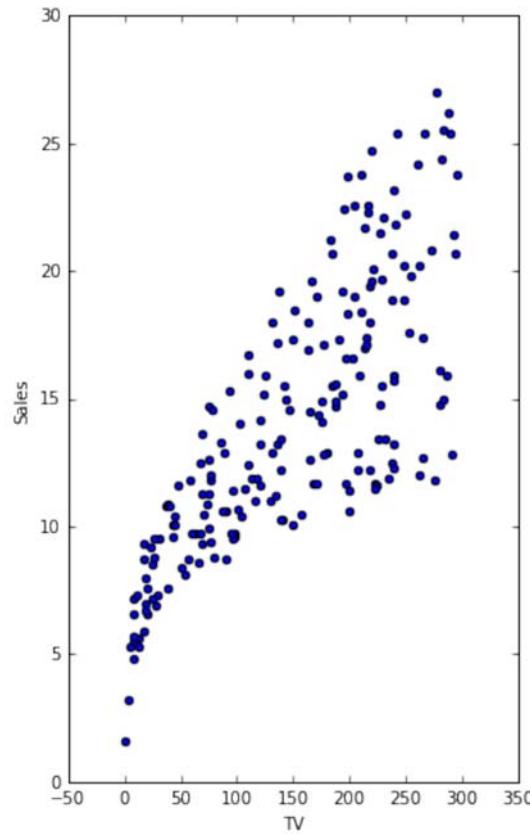
Aprendizaje supervisado

- ▶ Cuando la clase es de tipo nominal hablamos de clasificación.
 - ▶ En ocasiones, el numero de observaciones o ejemplos perteneciente a una de las clases es mucho mayor que el número de ejemplos perteneciente a otras, cuando esto pasa se habla de problema desequilibrado.
- ▶ Cuando la clase es de tipo numérico hablamos de regresión.

Minería de datos



Regresión



Algoritmos de
construcción de
ensembles y conjuntos
desequilibrados

José Francisco Díez
Pastor

Topics

10 Introducción a la
minería de datos

Ensembles para
problemas
desequilibrados

Contribuciones al
aprendizaje de
problemas
desequilibrados

Aplicaciones

Resumen y líneas
futuras

Término independiente 2.938

Coeficientes 'TV', 0.045, 'Radio', 0.188, 'Newspaper', -0.001



Algoritmos de
construcción de
ensembles y conjuntos
desequilibrados

José Francisco Díez
Pastor

Topics

11 Introducción a la
minería de datos

Ensembles para
problemas
desequilibrados

Contribuciones al
aprendizaje de
problemas
desequilibrados

Aplicaciones

Resumen y líneas
futuras

Clasificación. Árbol de decisión

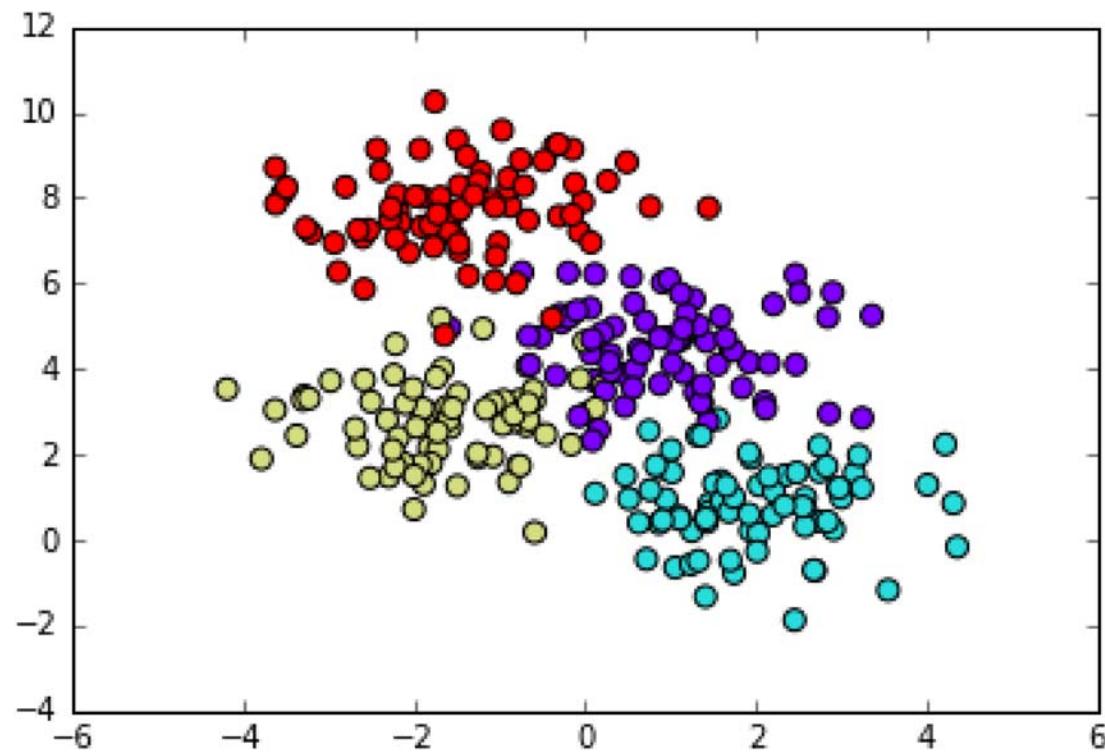
`buildDecisionTree(ejemplos):`

- ▶ Si todos los ejemplos son de la misma clase o número de ejemplos $\leq \text{minEjemplosHoja}$:
 - ▶ Hacer una hoja con ejemplos.
- ▶ Si no:
 - ▶ Best-Atr = atributo que mejor divide los ejemplos.
 - ▶ Se parte ejemplos en ejemplos1 y ejemplos2 usando Best-Atr.
 - ▶ `buildDecisionTree(ejemplos1)`
 - ▶ `buildDecisionTree(ejemplos2)`

Minería de datos



¿Qué es la minería de datos?



Algoritmos de
construcción de
ensembles y conjuntos
desequilibrados

José Francisco Díez
Pastor

Topics

12
Introducción a la
minería de datos

Ensembles para
problemas
desequilibrados

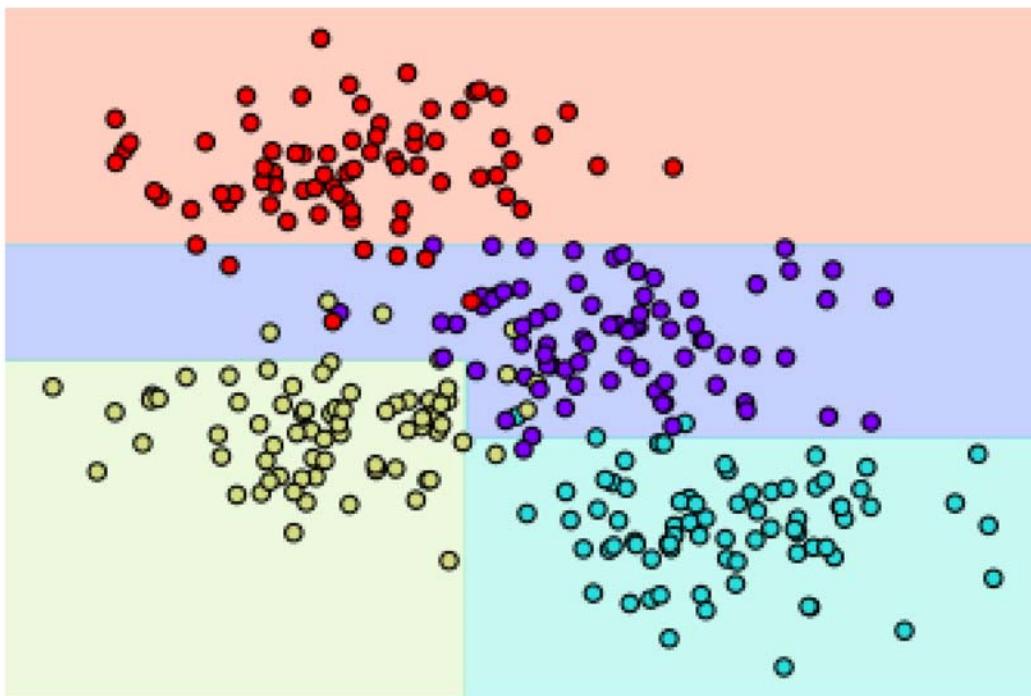
Contribuciones al
aprendizaje de
problemas
desequilibrados

Aplicaciones

Resumen y líneas
futuras



Clasificación. Árbol de decisión



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

13 Introducción a la minería de datos

Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

Ensembles



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

14 Introducción a la minería de datos

Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

- ▶ Los ensembles son combinaciones de múltiples clasificadores, a menudo llamados clasificadores base.
- ▶ La combinación de clasificadores a menudo ofrece mejores resultados que cualquiera de los clasificadores que forman el ensemble.

Ensembles



¿Porque funcionan?



- ▶ Idea intuitiva: Las decisiones difíciles se toman por un comité de expertos en lugar de solo uno.

Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

15 Introducción a la minería de datos

Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

Ensembles



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

16 Introducción a la minería de datos

Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

- ▶ El error de un ensemble depende del error de los clasificadores base y de la diversidad entre los distintos clasificadores base. La diversidad es clave para el buen funcionamiento de los ensembles.
- ▶ No tendría sentido un comité de expertos si todos ellos tienen la misma opinión, cometan los mismos errores o son expertos en el mismo área.



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

17 Introducción a la minería de datos

Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

Bagging y Boosting

- ▶ Son las dos técnicas más usadas para crear ensembles.
- ▶ En Bagging cada uno de los N clasificadores base se construye con una versión del conjunto de datos creada mediante remuestreo con remplazamiento.
- ▶ En Boosting el clasificador base N se construye con un remuestreo con remplazamiento del conjunto original donde se da más peso a las instancias falladas por los clasificadores anteriores.



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

18 Introducción a la minería de datos

Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

Bagging y Boosting

- ▶ Ejemplo: Realizar un cuestionario online en grupo. Hay 10 temas.
 - ▶ Bagging: Cada alumno se estudia al azar 10 temas, algunos alumnos pueden estudiar el mismo tema dos veces. A la hora de contestar votan.
 - ▶ Boosting: El alumno 1 estudia 10 temas al azar, hace autoevaluación y asigna un peso mayor a los temas en los que ha fallado. El alumno 2 tiene en cuenta los pesos a la hora de elegir los temas al azar etc. A la hora de contestar votan, pero los alumnos con mejor autoevaluación tienen más peso.



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

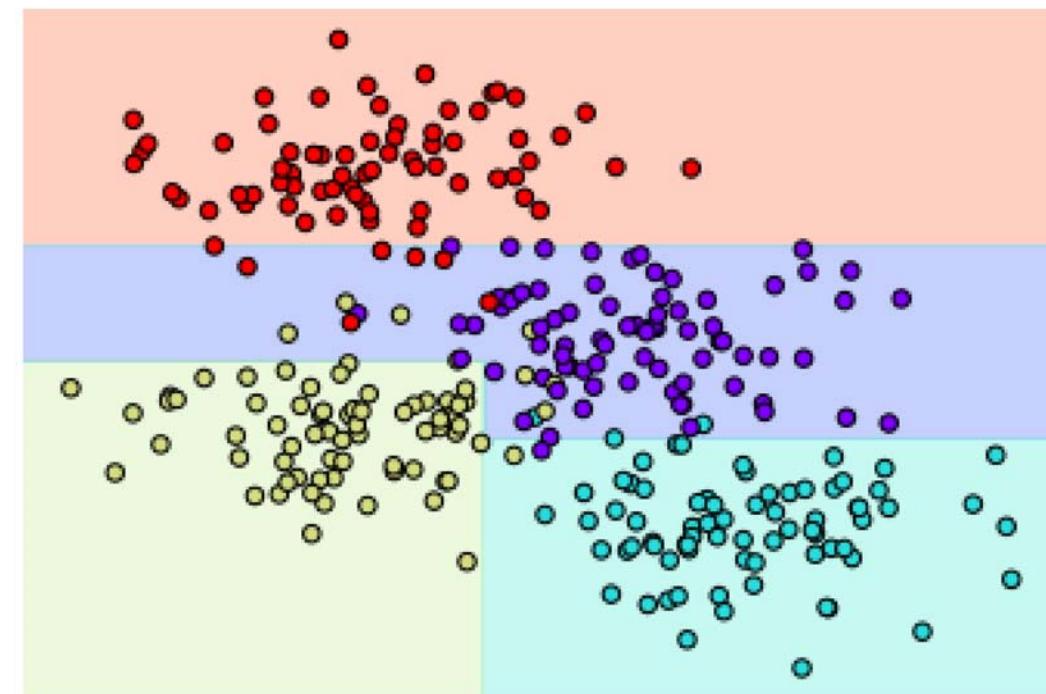
19 Introducción a la minería de datos

Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras





Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

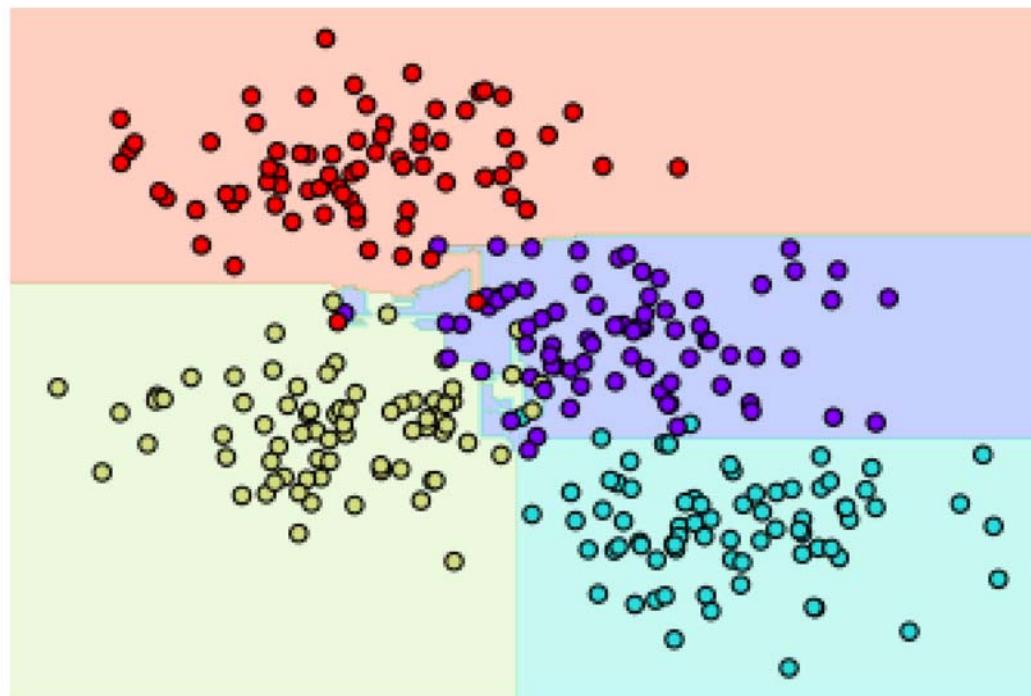
20 Introducción a la minería de datos

Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras



Ensembles



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

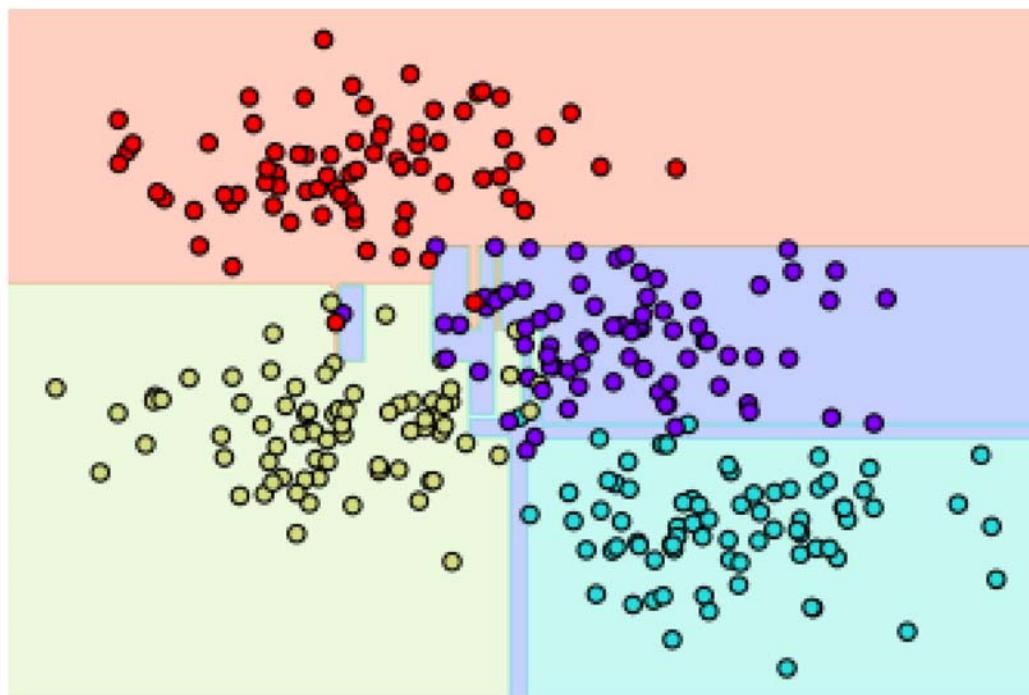
21 Introducción a la minería de datos

Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras





Algoritmos de
construcción de
ensembles y conjuntos
desequilibrados

José Francisco Díez
Pastor

Topics

Introducción a la
minería de datos

22 Ensembles para
problemas
desequilibrados

Contribuciones al
aprendizaje de
problemas
desequilibrados

Aplicaciones

Resumen y líneas
futuras

Ensembles para problemas desequilibrados

Introducción al aprendizaje desequilibrado



Algoritmos de
construcción de
ensembles y conjuntos
desequilibrados

José Francisco Díez
Pastor

Topics

Introducción a la
minería de datos

23
Ensembles para
problemas
desequilibrados

Contribuciones al
aprendizaje de
problemas
desequilibrados

Aplicaciones

Resumen y líneas
futuras

Los problemas desequilibrados ocurren cuando tenemos muchas más instancias de una clase que de otra.

Este tipo de problemas son comunes en medicina, seguridad, ingeniería, finanzas, etc.

Introducción al aprendizaje desequilibrado



Son necesarios métodos específicos para desequilibrados por 3 razones:

1. Los clasificadores estandar maximizan el porcentaje de acierto, luego la clase minoritaria puede ser ignorada.
2. Los clasificadores estandar suponen que el conjunto de entrenamiento es una representación fiel del problema.
3. Los errores en problemas desequilibrados pueden tener distintos costes en función de la clase.

Además de la perdida de rendimiento debido al desbalanceo, los conjuntos desequilibrados presentan otros problemas como: solape entre clases, falta de densidad, ruido, ejemplos borderline, etc.

- S. Visa, A. Ralescu, Issues in mining imbalanced data sets - a review paper, in: Proceedings of the Sixteen Midwest Artificial Intelligence and Cognitive Science Conference, 2005, pp. 67–73.

Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

24
Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

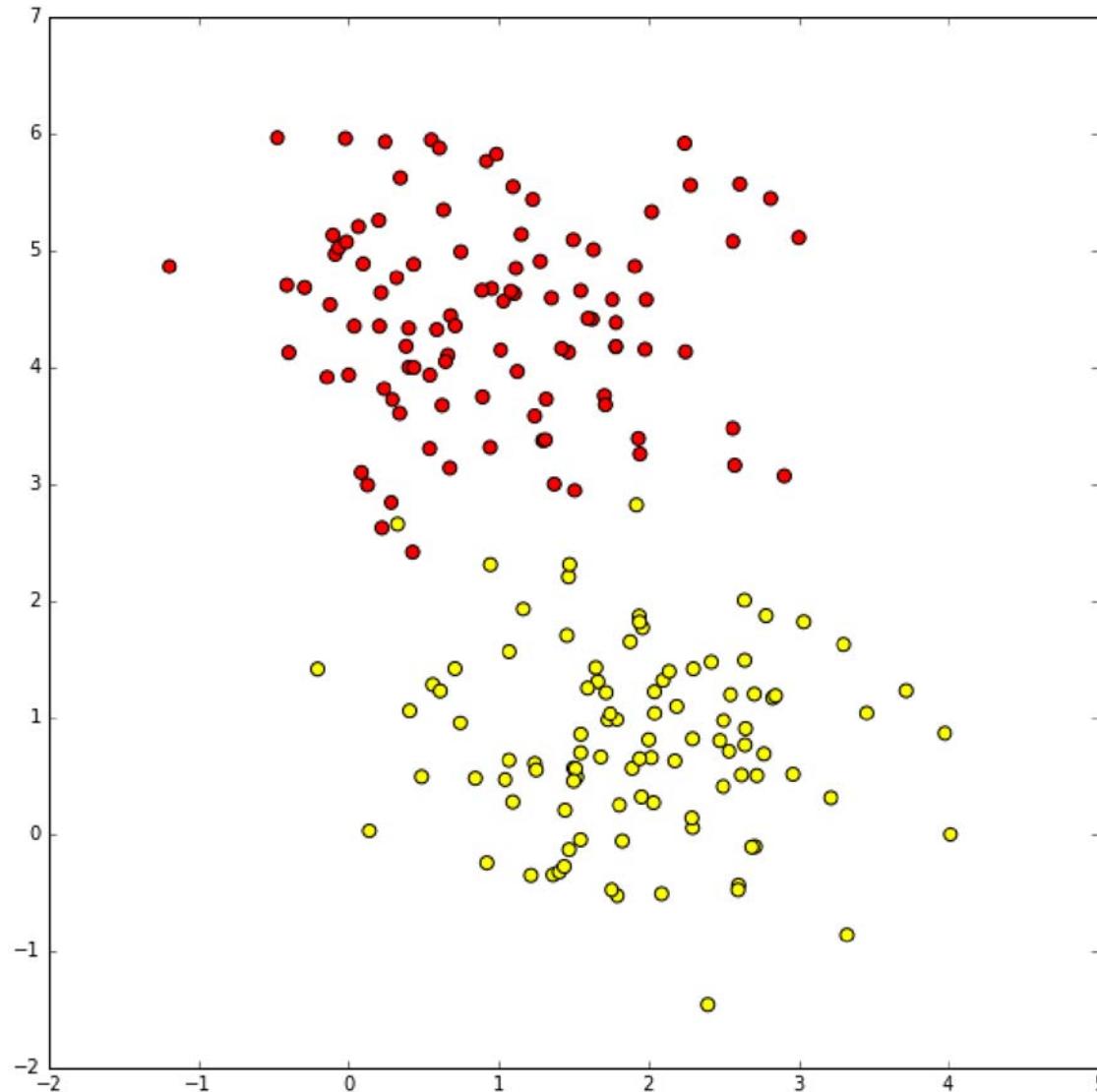
Aplicaciones

Resumen y líneas futuras

Introducción al aprendizaje desequilibrado



Ratio = 1



Algoritmos de
construcción de
ensembles y conjuntos
desequilibrados

José Francisco Díez
Pastor

Topics

Introducción a la
minería de datos

25 Ensembles para
problemas
desequilibrados

Contribuciones al
aprendizaje de
problemas
desequilibrados

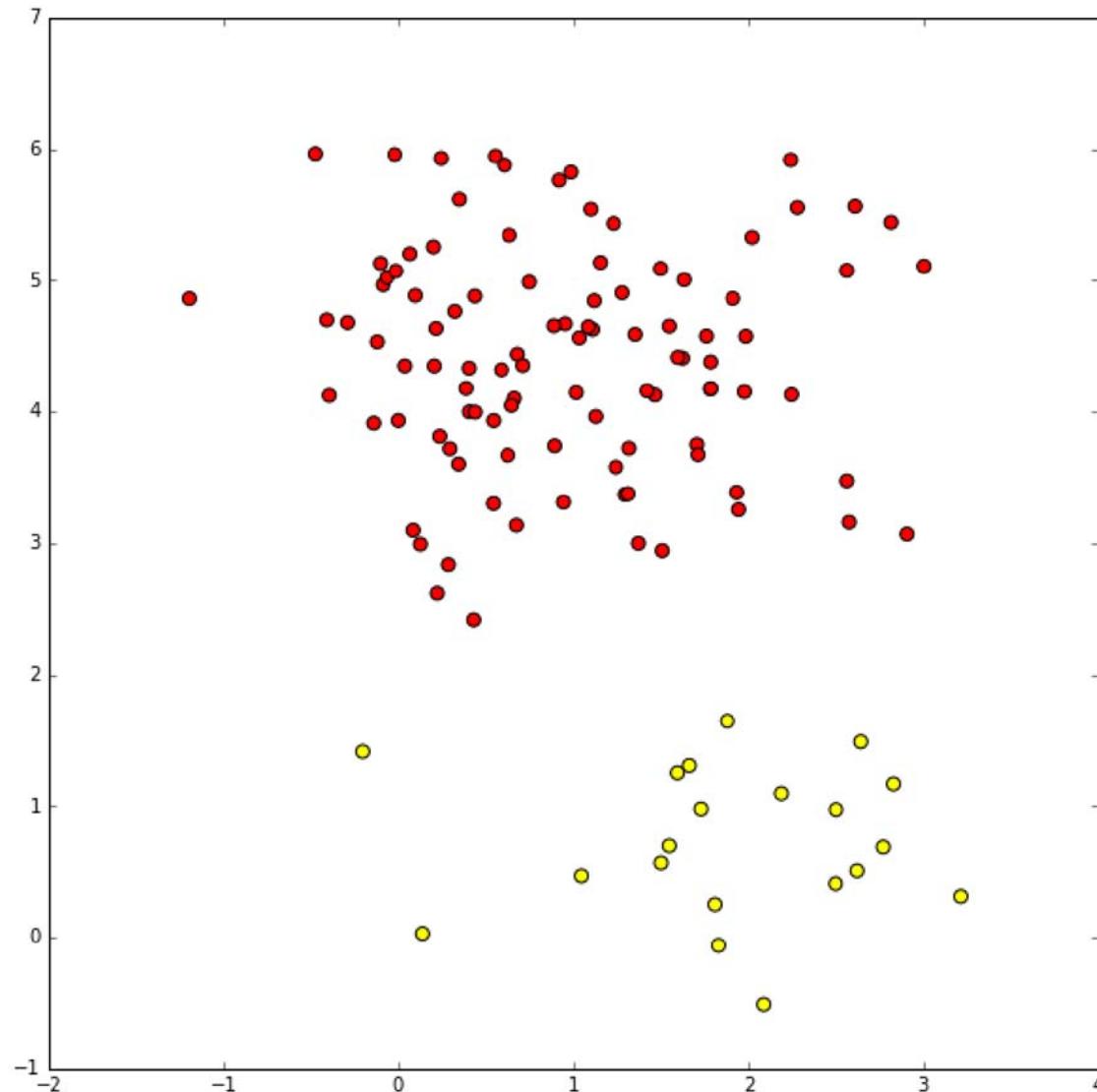
Aplicaciones

Resumen y líneas
futuras

Introducción al aprendizaje desequilibrado



Ratio = 0.2



Algoritmos de
construcción de
ensembles y conjuntos
desequilibrados

José Francisco Díez
Pastor

Topics

Introducción a la
minería de datos

26
Ensembles para
problemas
desequilibrados

Contribuciones al
aprendizaje de
problemas
desequilibrados

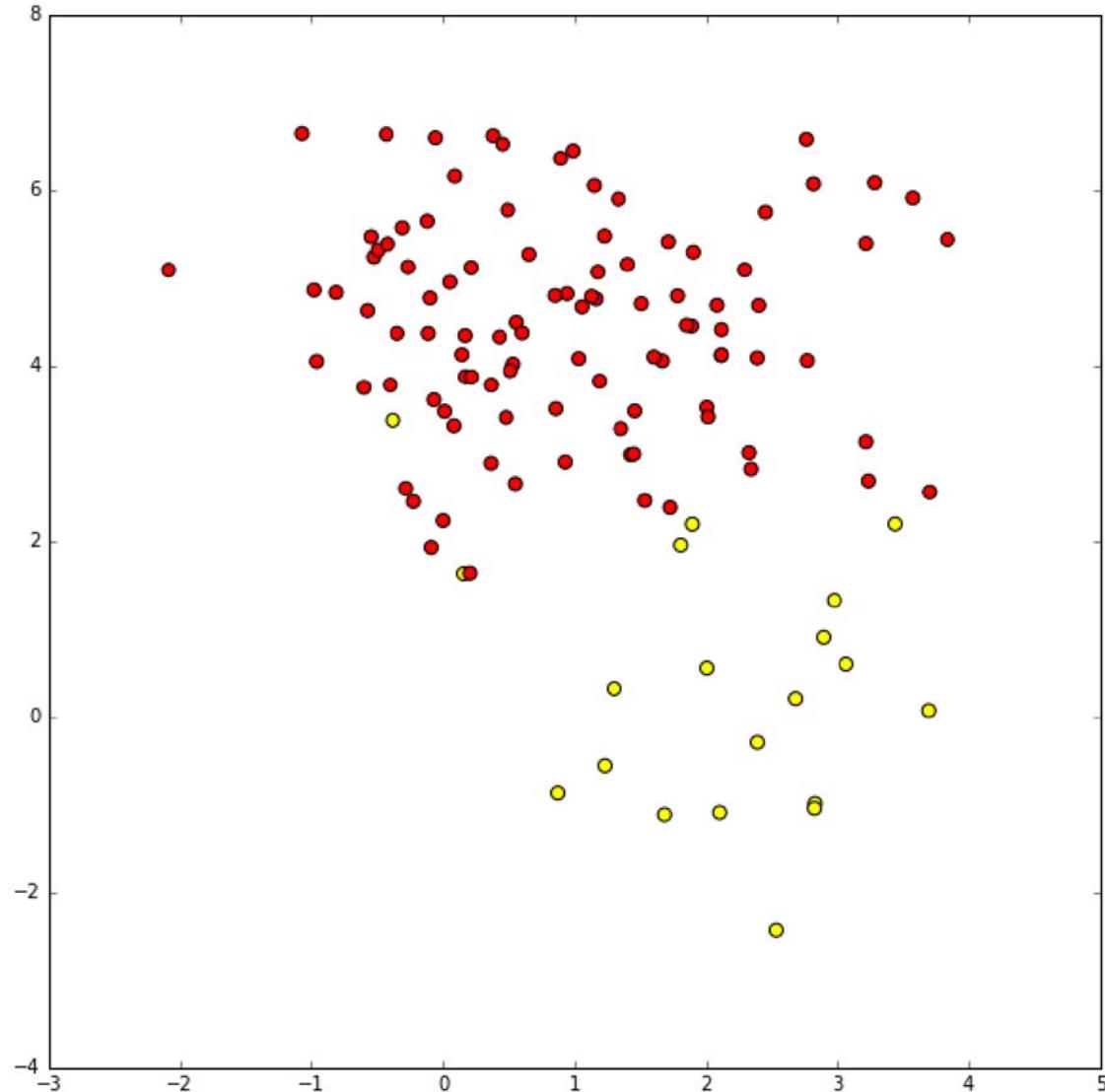
Aplicaciones

Resumen y líneas
futuras

Introducción al aprendizaje desequilibrado



Ratio = 0.2 (Solape)



Algoritmos de
construcción de
ensembles y conjuntos
desequilibrados

José Francisco Díez
Pastor

Topics

Introducción a la
minería de datos

27
Ensembles para
problemas
desequilibrados

Contribuciones al
aprendizaje de
problemas
desequilibrados

Aplicaciones

Resumen y líneas
futuras

Introducción al aprendizaje desequilibrado



De acuerdo a Galar et al. técnicas para lidiar con conjuntos desequilibrados se pueden clasificar en 4 niveles:

1. Nivel de algoritmo.
2. Nivel de datos.
3. Cost-sensitive.
4. Ensemble learning.

En general los niveles de algoritmo y cost-sensitive son más dependientes de los datos y el nivel de datos y ensemble son más versátiles.

- M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, F. Herrera, A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches, *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 42 (4) (2012) 463 –484.

Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

28
Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

Nivel de datos: Técnicas de preprocesado



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez
Pastor

Topics

Introducción a la minería de datos

29
Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

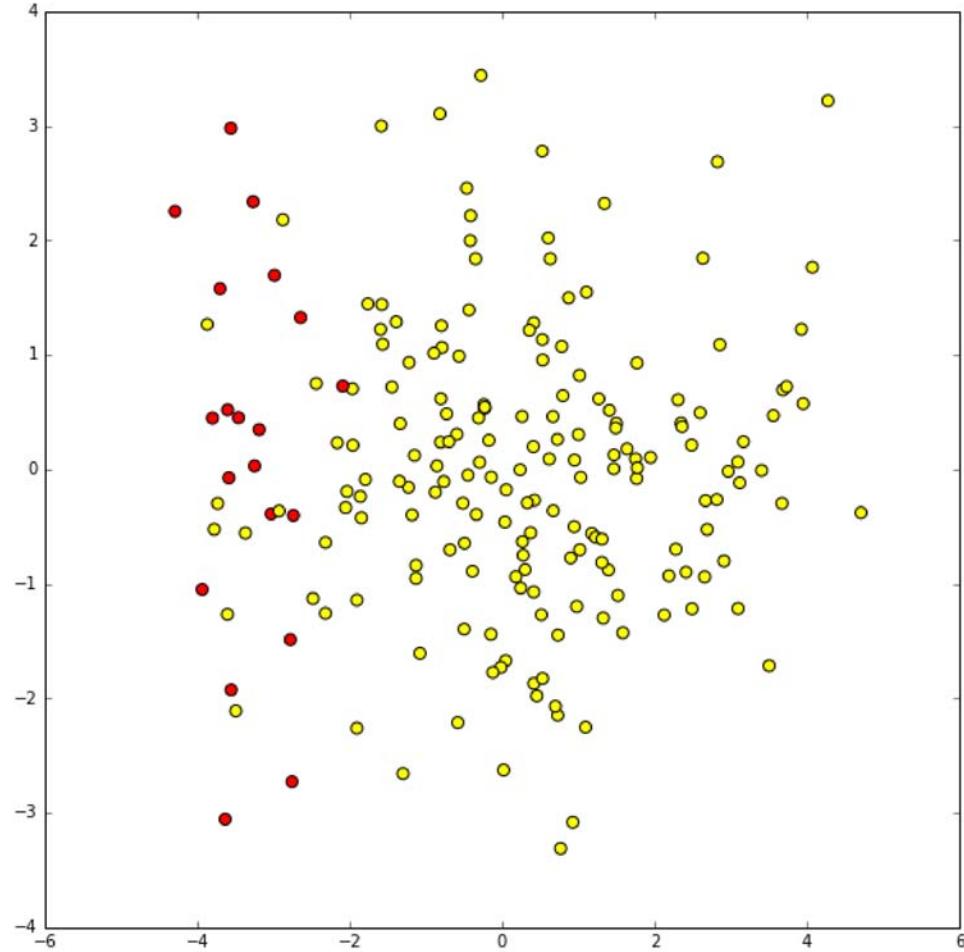
Las técnicas de preprocesado tratan de equilibrar el ratio entre las clases. Las estrategias más comunes reducen el tamaño de la clase mayoritaria, reducen el de la minoritaria o ambas cosas al mismo tiempo.

- ▶ Random Undersampling.
- ▶ Random Oversampling.
- ▶ SMOTE (Synthetic Minority Over-sampling Technique).

Nivel de datos: Técnicas de preprocesado



Random Undersampling (Antes)



Consiste en eliminar aleatoriamente algunos ejemplos de la clase mayoritaria.

Algoritmos de construcción de ensambles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

30
Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

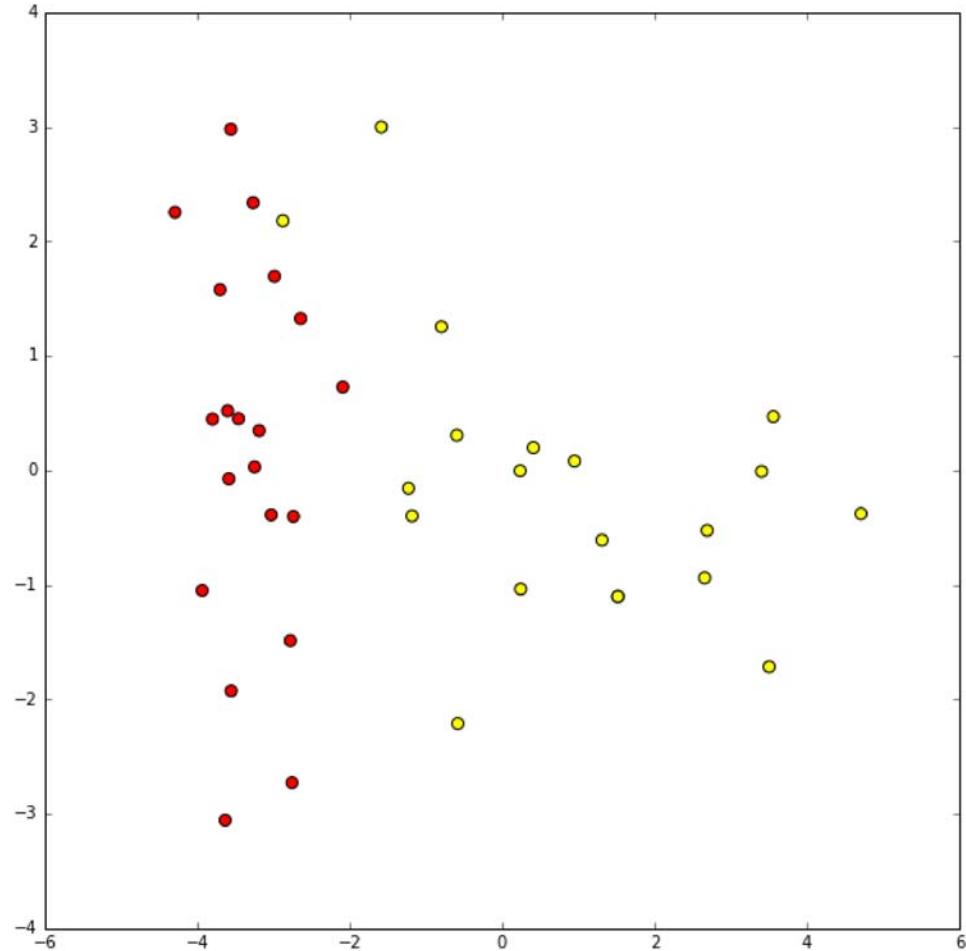
Aplicaciones

Resumen y líneas futuras



Nivel de datos: Técnicas de preprocesado

Random Undersampling (Después)



Consiste en eliminar aleatoriamente algunos ejemplos de la clase mayoritaria.

Algoritmos de construcción de ensambles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

31
Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

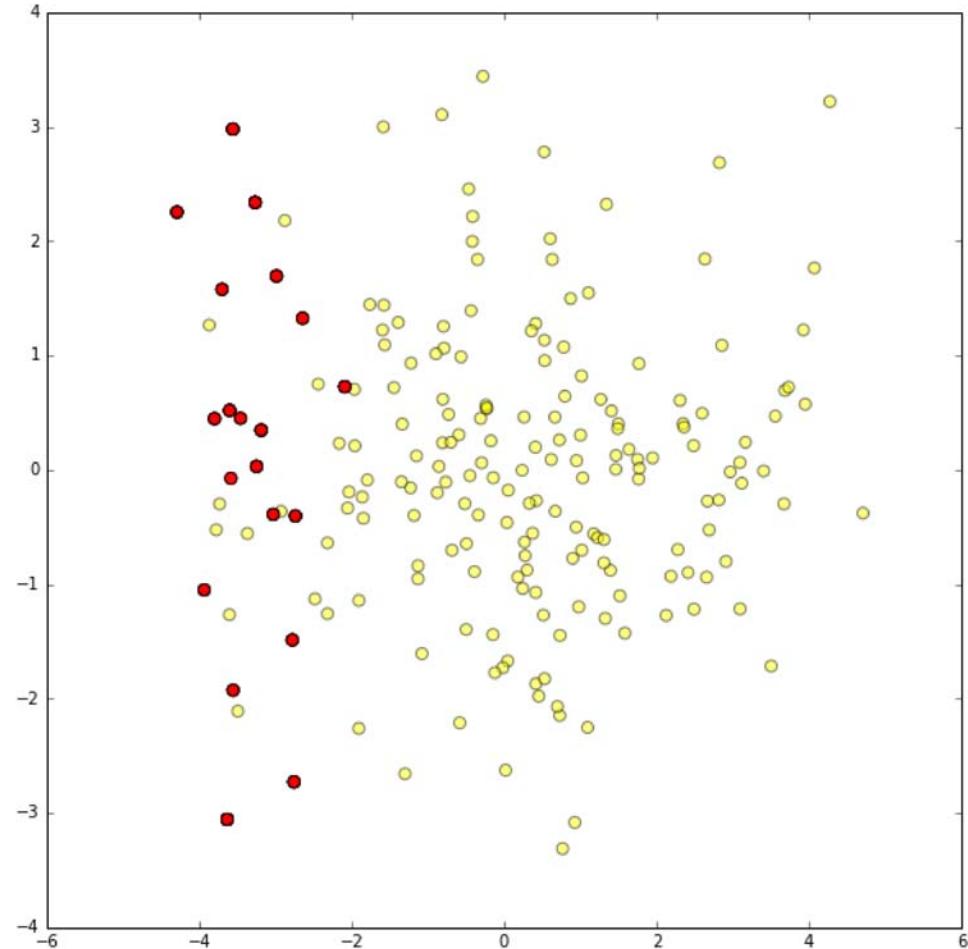
Aplicaciones

Resumen y líneas futuras

Nivel de datos: Técnicas de preprocesado



Random Oversampling (Después)



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

32
Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

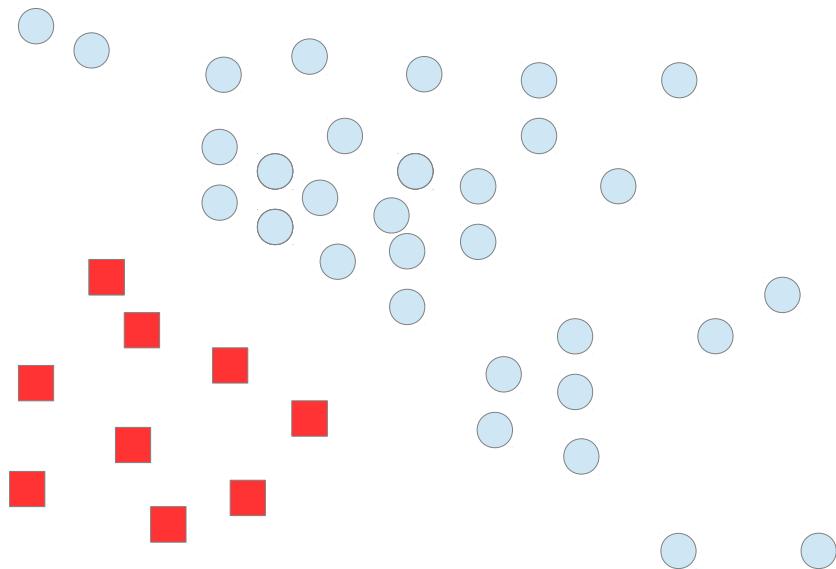
Resumen y líneas futuras

Consiste en añadir copias exactas de algunos de los ejemplos de la clase minoritaria.

Nivel de datos: Técnicas de preprocesado



SMOTE



SMOTE (Synthetic Minority Over-sampling Technique), en lugar de crear copias, crea instancias artificiales.

Algoritmos de construcción de ensambles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

33

Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

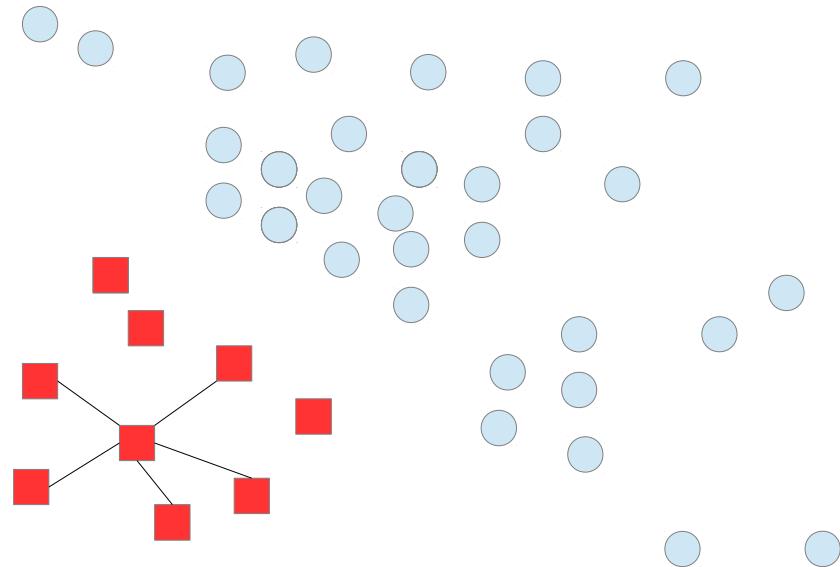
Resumen y líneas futuras

86

Nivel de datos: Técnicas de preprocesado



SMOTE



Para cada instancia uno de sus k vecinos más cercanos (de su misma clase) es elegido aleatoriamente.

Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

34

Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

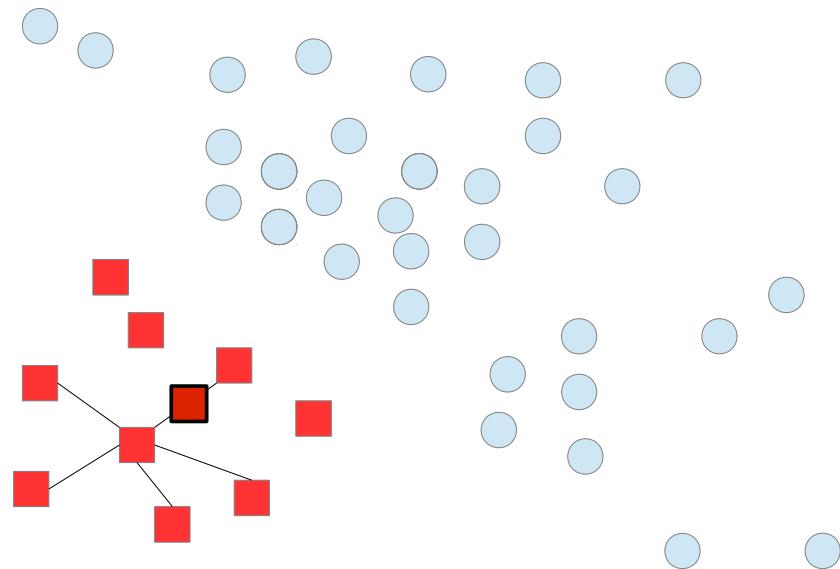
Aplicaciones

Resumen y líneas futuras

Nivel de datos: Técnicas de preprocesado



SMOTE



Entonces, un nuevo ejemplo es creado en un punto aleatorio en el segmento definido entre la instancia y su vecino.

Algoritmos de construcción de ensambles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

35
Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

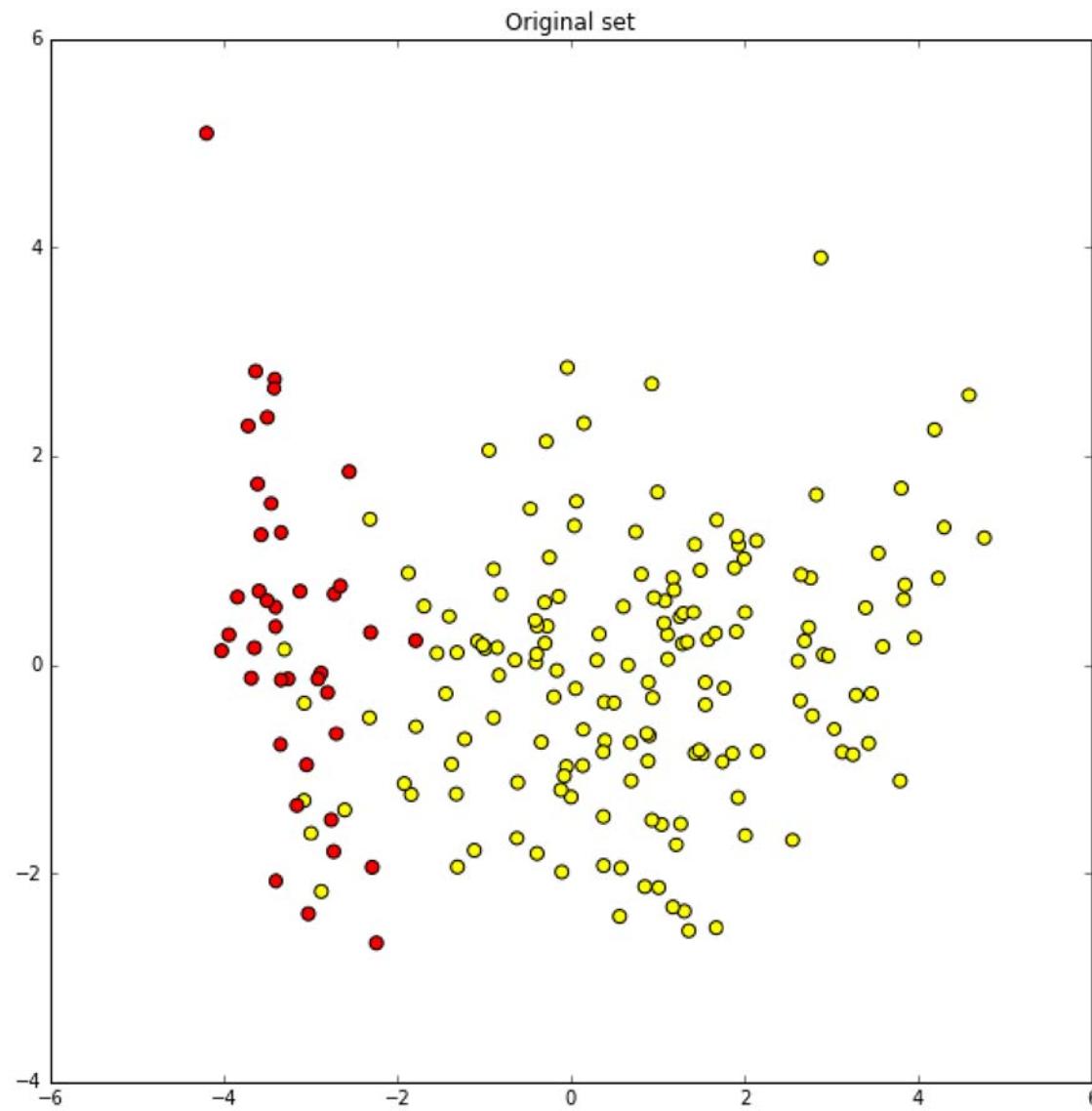
Aplicaciones

Resumen y líneas futuras



Nivel de datos: Técnicas de preprocesado

SMOTE (Antes)



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

36
Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

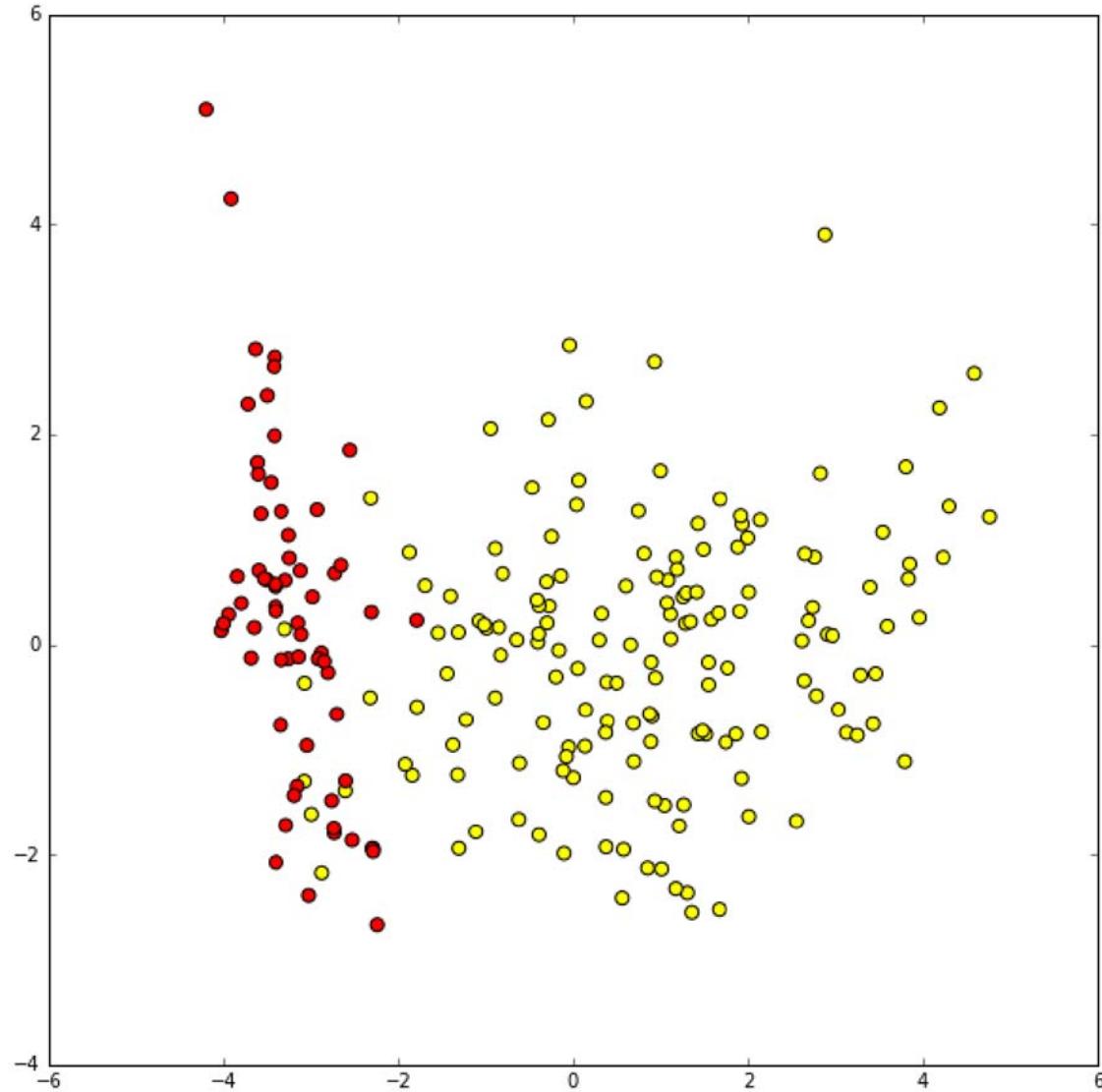
Aplicaciones

Resumen y líneas futuras

Nivel de datos: Técnicas de preprocesado



SMOTE 100% (Después)



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

37
Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

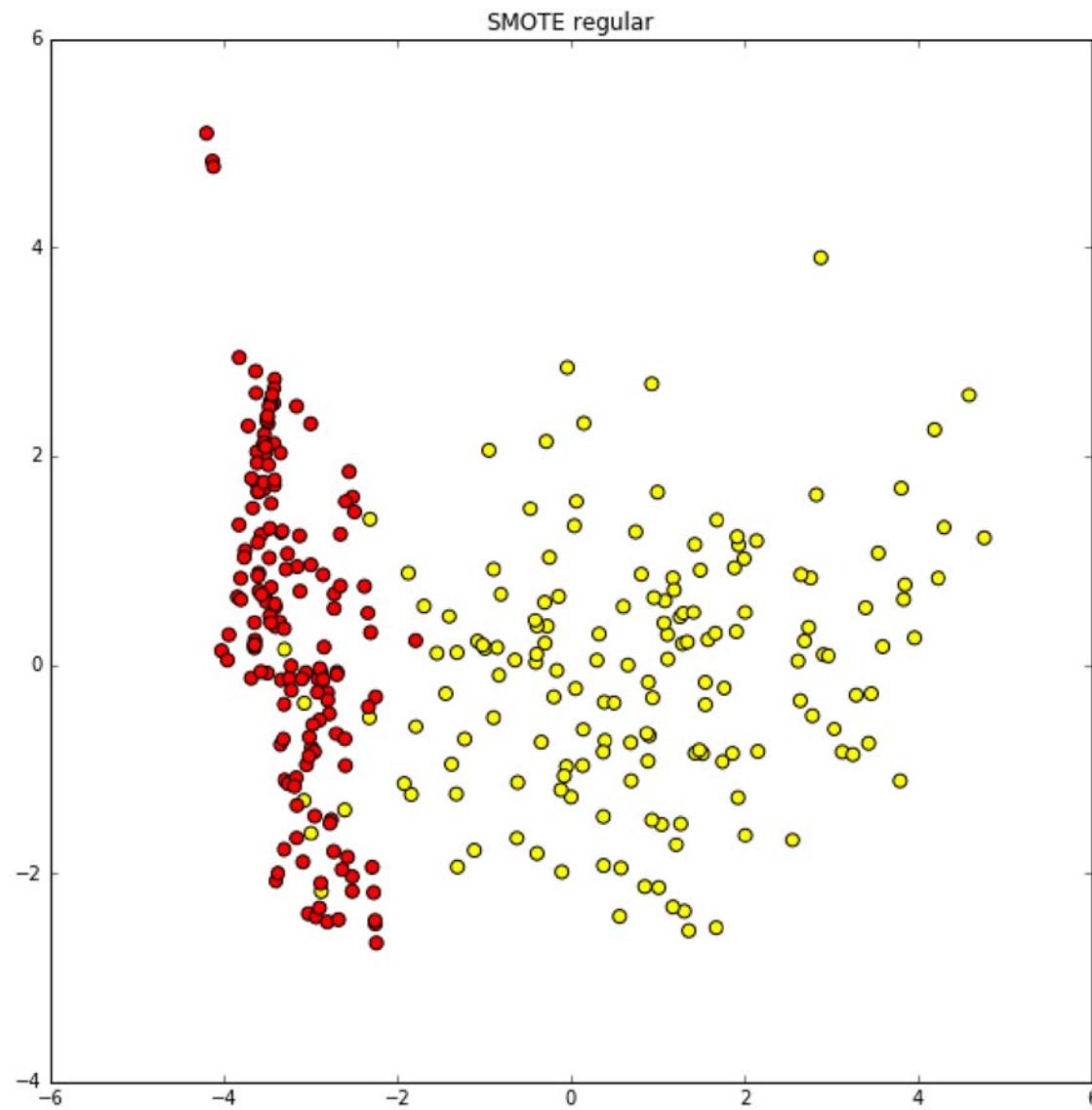
Aplicaciones

Resumen y líneas futuras

Nivel de datos: Técnicas de preprocesado



SMOTE hasta igualar (Después)



Algoritmos de
construcción de
ensembles y conjuntos
desequilibrados

José Francisco Díez
Pastor

Topics

Introducción a la
minería de datos

38
Ensembles para
problemas
desequilibrados

Contribuciones al
aprendizaje de
problemas
desequilibrados

Aplicaciones

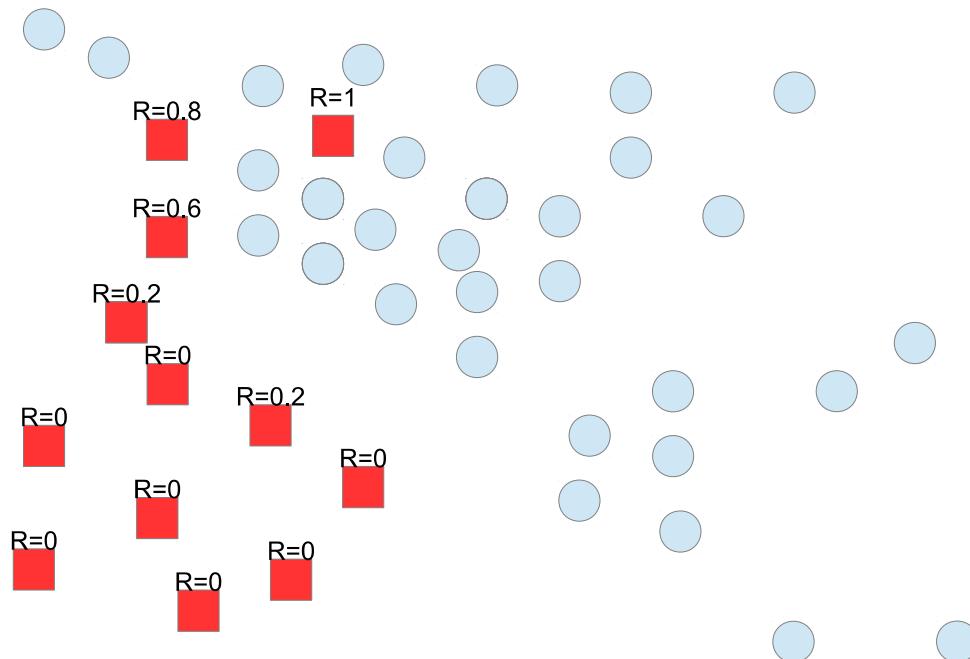
Resumen y líneas
futuras



Nivel de datos: Técnicas de preprocesado

ADASYN

$R = \text{Num vecinos mayoritaria} / \text{num vecinos}$



ADASYN es una técnica similar a SMOTE, pero en lugar de crear instancias artificiales de manera uniforme lo hace de acuerdo a unos pesos calculados en función de la dificultad para predecir las instancias.

Algoritmos de construcción de ensambles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

39
Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

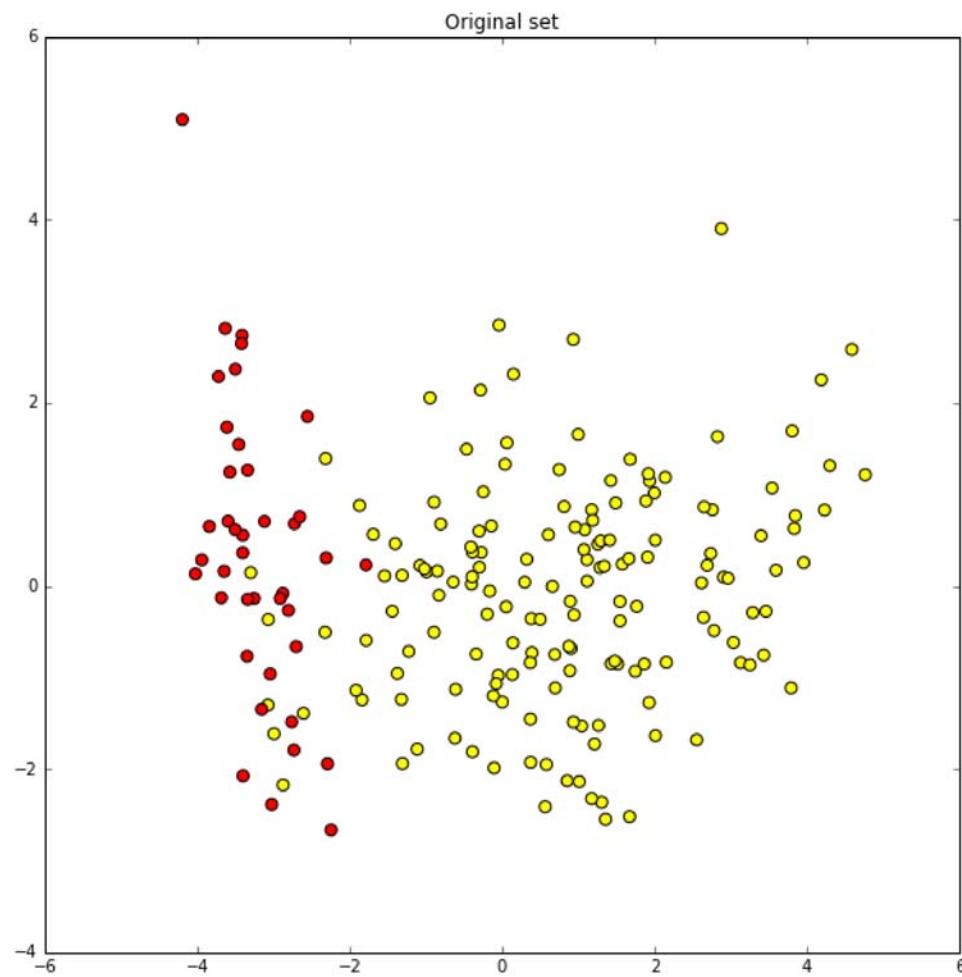
Aplicaciones

Resumen y líneas futuras

Nivel de datos: Técnicas de preprocesado



ADASYN (Antes)



Algoritmos de
construcción de
ensembles y conjuntos
desequilibrados

José Francisco Díez
Pastor

Topics

Introducción a la
minería de datos

40
Ensembles para
problemas
desequilibrados

Contribuciones al
aprendizaje de
problemas
desequilibrados

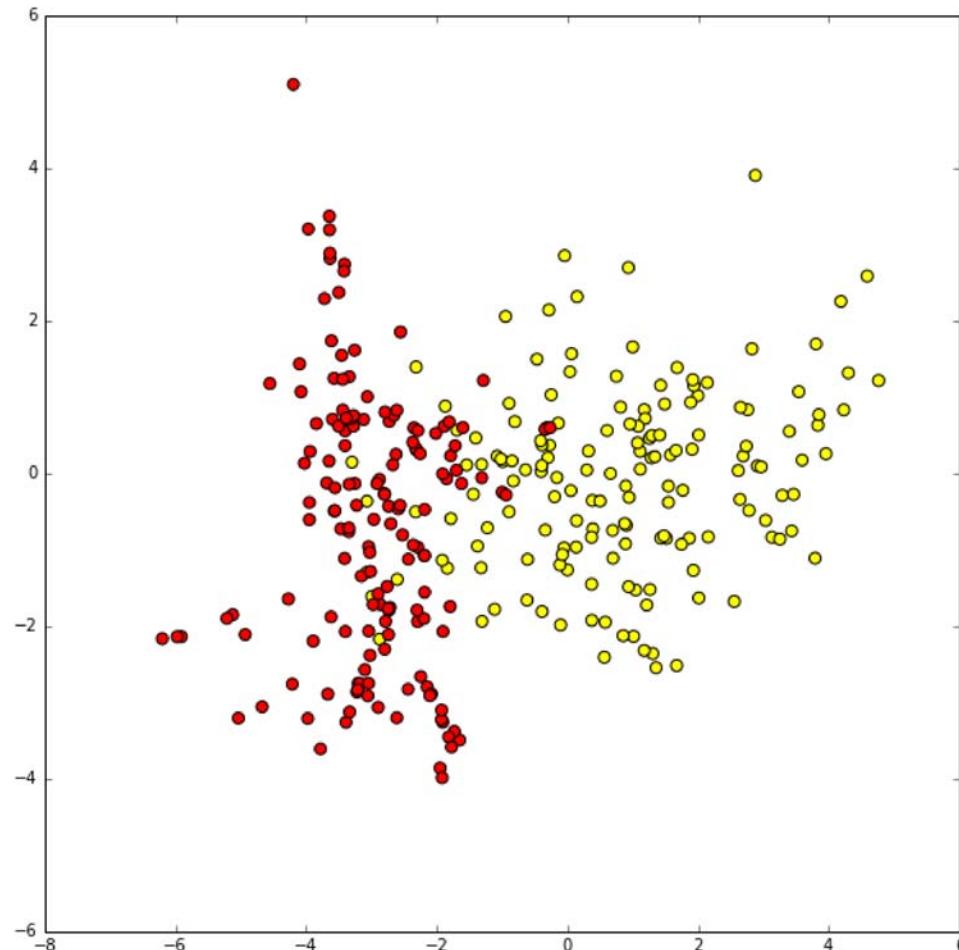
Aplicaciones

Resumen y líneas
futuras

Nivel de datos: Técnicas de preprocesado



ADASYN (Después)



Las instancias más cercanas a la frontera tendrán más instancias artificiales en su vecindad. Haciendo que sean más fáciles de clasificar.

Algoritmos de construcción de ensambles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

41
Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

Ensemble learning



Las técnicas de preprocessamiento aumentan el rendimiento individual y como tienen una componente aleatoria, también incrementan la diversidad del ensemble.

Hay dos estrategias para construir ensembles para conjuntos desequilibrados:

- ▶ Usar diferentes conjuntos preprocessados para cada uno de los clasificadores base.
- ▶ Combinar técnicas de ensembles tradicionales con técnicas de preprocessamiento: SMOTEBagging, SMOTEBoost y RUSBoost.

Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

42

Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

Ensemble learning.



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

43
Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

SMOTEBagging

- ▶ Similar a Bagging. Cada clasificador base se entrena con un dataset S_i cuyas clases tienen el mismo tamaño.
- ▶ S_i está formado por un remuestreo de la mayoritaria y una combinación de Oversampling y SMOTE de la minoritaria.

(Ver dibujo)

Ensemble learning.



Algoritmos de construcción de ensambles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

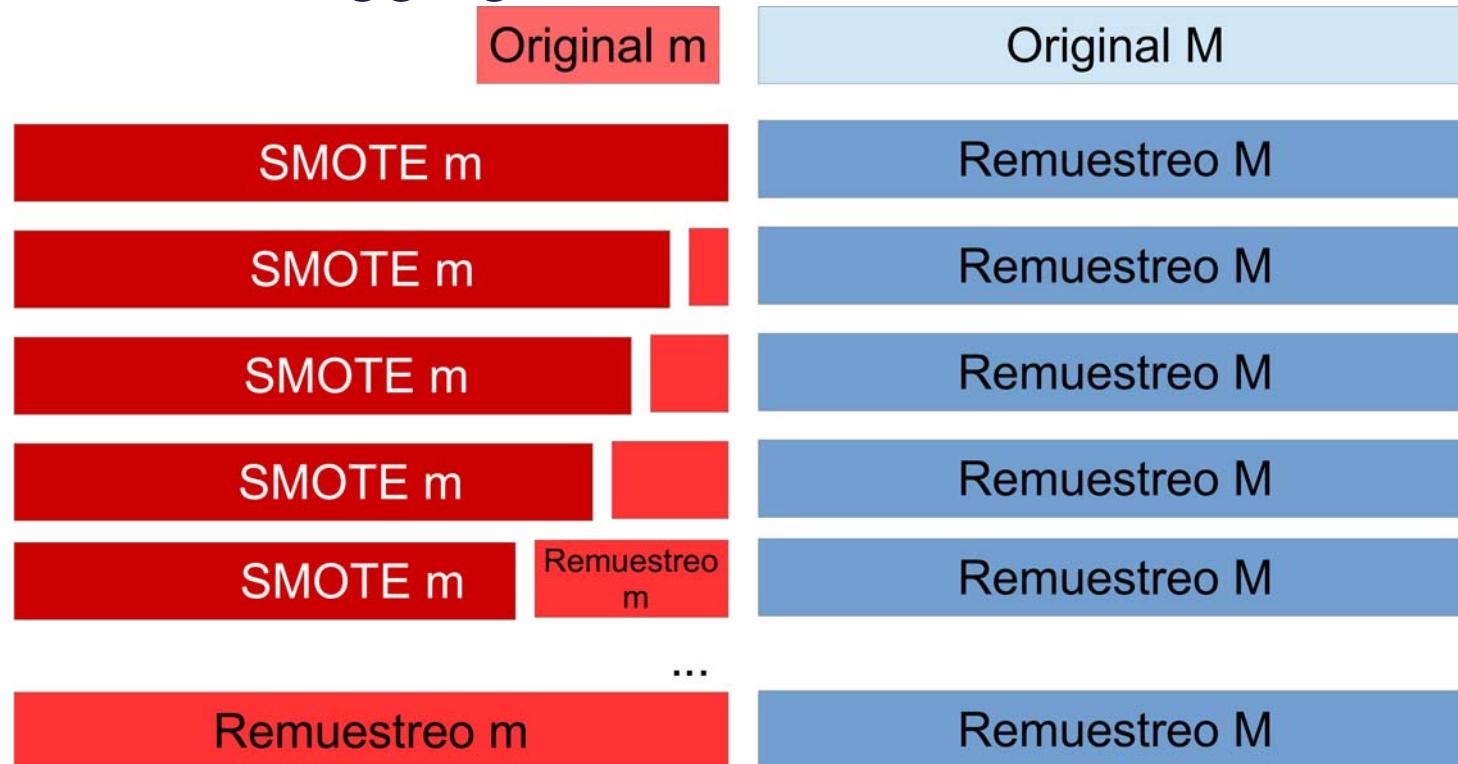
44
Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

SMOTEBagging



Ensemble learning.



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

45

Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

RUSBoost

- ▶ Una modificación de AdaBoost.
- ▶ Después de cada ronda, se aplica Random Undersampling hasta igualar el tamaño de las clases.
- ▶ Las instancias originales tienen pesos de acuerdo a una pseudoperdida. Estos pesos se normalizan para que la suma de pesos siga siendo igual al tamaño del conjunto de datos.

86



Ensemble learning.

RUSBoost

Algorithm RUSBoost

Given: Set S of examples $(x_1, y_1), \dots, (x_m, y_m)$ with minority class $y^r \in Y$, $|Y| = 2$

Weak learner, $WeakLearn$

Number of iterations, T

Desired percentage of total instances to be represented by the minority class, N

- 1 Initialize $D_1(i) = \frac{1}{m}$ for all i .
- 2 Do for $t = 1, 2, \dots, T$
 - a Create temporary training dataset S'_t with distribution D'_t using random undersampling
 - b Call $WeakLearn$, providing it with examples S'_t and their weights D'_t .
 - c Get back a hypothesis $h_t : X \times Y \rightarrow [0, 1]$.
 - d Calculate the pseudo-loss (for S and D_t):
$$\epsilon_t = \sum_{(i,y):y_i \neq y} D_t(i)(1 - h_t(x_i, y_i) + h_t(x_i, y)).$$
 - e Calculate the weight update parameter:
$$\alpha_t = \frac{\epsilon_t}{1 - \epsilon_t}.$$
 - f Update D_t :
$$D_{t+1}(i) = D_t(i) \alpha_t^{\frac{1}{2}(1+h_t(x_i,y_i)-h_t(x_i,y:y\neq y_i))}.$$
 - g Normalize D_{t+1} : Let $Z_t = \sum_i D_{t+1}(i)$.
$$D_{t+1}(i) = \frac{D_{t+1}(i)}{Z_t}.$$
- 3 Output the final hypothesis:
$$H(x) = \operatorname{argmax}_{y \in Y} \sum_{t=1}^T h_t(x, y) \log \frac{1}{\alpha_t}.$$

Algoritmos de construcción de ensambles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

46
Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

Ensemble learning.



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

47

Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

SMOTEBoost

- ▶ Una modificación de AdaBoost.
- ▶ Después de cada ronda, se aplica SMOTE para crear instancias artificiales de la minoritaria.
- ▶ Las instancias artificiales tienen peso uniforme. Las instancias originales tienen pesos de acuerdo a una pseudoperdida. Las instancias difíciles para los clasificadores anteriores tienen mayor peso.

86

Ensemble learning.



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

48 Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

RAMOBost

- ▶ Puede verse como una versión de SMOTEBoost que utiliza ADASYN en lugar de SMOTE.

Medidas para conjuntos desequilibrados



En problemas binarios desequilibrados, las instancias pueden ser etiquetadas como positivas p (minoritaria) o negativas n (mayoritaria). Para cada predicción hay 4 posibles opciones:

Table : Matriz de confusión en problemas binarios

	Positive prediction	Negative prediction
Positive class	True Positive (TP)	False Negative (FN)
Negative class	False Positive (FP)	True Negative (TN)

- ▶ Recall o True Positive Rate (TPR) se define como TP/P .
- ▶ False Positive Rate (FPR) se define como FP/N .
- ▶ Precision se define como $TP/(TP + FP)$.

Algoritmos de construcción de ensambles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

49
Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

Medidas para conjuntos desequilibrados



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

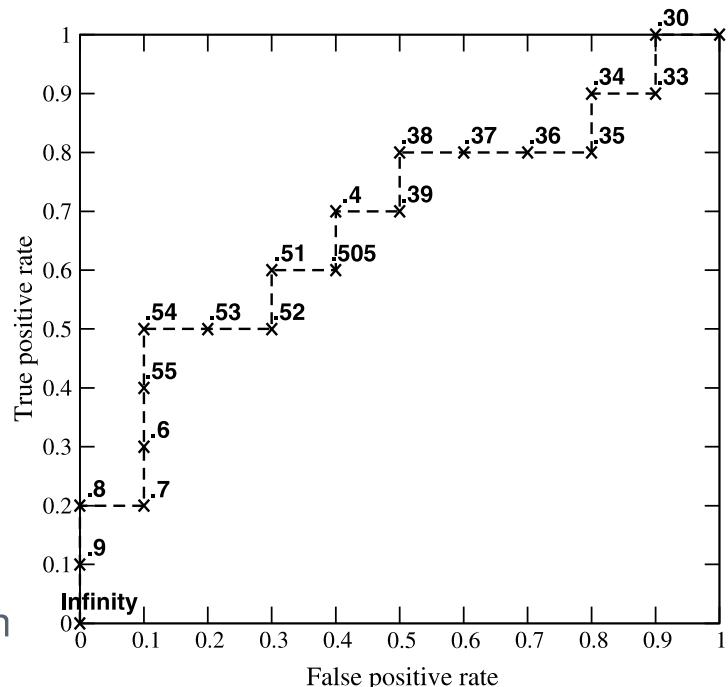
Introducción a la minería de datos

50
Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras



La curva ROC es una representación visual 2D del rendimiento de un clasificador. El área bajo la curva ROC (AUC) representa el rendimiento de un clasificador binario con un escalar.

$$\text{F-Measure} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

$$\text{Geometric Mean} = \sqrt{\text{TP}/P \times \text{TN}/N}$$



Algoritmos de
construcción de
ensembles y conjuntos
desequilibrados

José Francisco Díez
Pastor

Topics

Introducción a la
minería de datos

Ensembles para
problemas
desequilibrados

51 Contribuciones al
aprendizaje de
problemas
desequilibrados

Aplicaciones

Resumen y líneas
futuras

Contribuciones al aprendizaje de problemas desequilibrados

Random Balance: Motivación



Algoritmos de
construcción de
ensembles y conjuntos
desequilibrados

José Francisco Díez
Pastor

Topics

Introducción a la
minería de datos

Ensembles para
problemas
desequilibrados

52 Contribuciones al
aprendizaje de
problemas
desequilibrados

Aplicaciones

Resumen y líneas
futuras

Al trabajar con desequilibrados hay principalmente
3 enfoques de nivel de datos:

- ▶ Incrementar el tamaño de la clase minoritaria usando SMOTE.
- ▶ Reducir el tamaño de la maoritaria usando Random Undersampling.
- ▶ Hacer ambas cosas al mismo tiempo: oversampling y undersampling.

Problema: La técnica y las proporciones óptimas
son dependientes del problema y difíciles de
encontrar.

Random Balance: Motivación



¿Es posible confiar completamente en la aleatoriedad y la repetición para superar el problema de encontrar la técnica y sus parámetros?

- ▶ **Díez-Pastor, J. F., Rodríguez, J. J., García-Osorio, C. & Kuncheva, L. (2015). Random Balance: Ensembles of Variable Priors Classifiers for Imbalanced Data. Knowledge-Based Systems. In press.**

Algoritmos de construcción de ensambles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

Ensembles para problemas desequilibrados

53 Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

Random Balance y RB-Boost



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez
Pastor

Topics

Introducción a la minería de datos

Ensembles para problemas desequilibrados

54 Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

- ▶ Presentamos una nueva técnica de procesamiento llamada Random Balance. Esta técnica puede ser usada dentro de un ensemble para incrementar la diversidad y tratar con el problema del desequilibrio.
- ▶ Además se describe una nuevo ensemble para conjuntos desequilibrados llamado RB-Boost (Random Balance Boost) el cual es una modificación con Random Balance de AdaBoost.M2.

Random Balance



Un conjunto de datos del mismo tamaño que el original es obtenido para cada uno de los clasificadores base del ensemble donde el ratio es elegido aleatoriamente.

- ▶ Iteración 1, el ratio de desequilibrio se reduce.
- ▶ Iteración 2, el ratio se invierte.
- ▶ Iteración 3, la minoritaria se vuelve incluso más pequeña.

Todos estos casos son posibles dado que el ratio es aleatorio.

Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

Ensembles para problemas desequilibrados

55 Contribuciones al aprendizaje de problemas desequilibrados

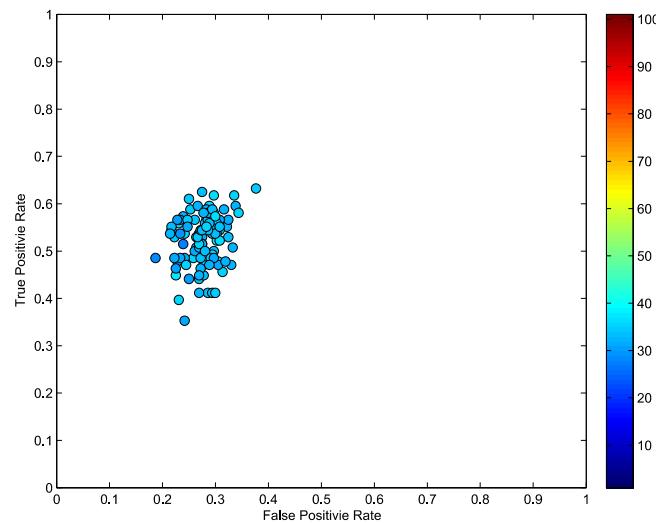
Aplicaciones

Resumen y líneas futuras

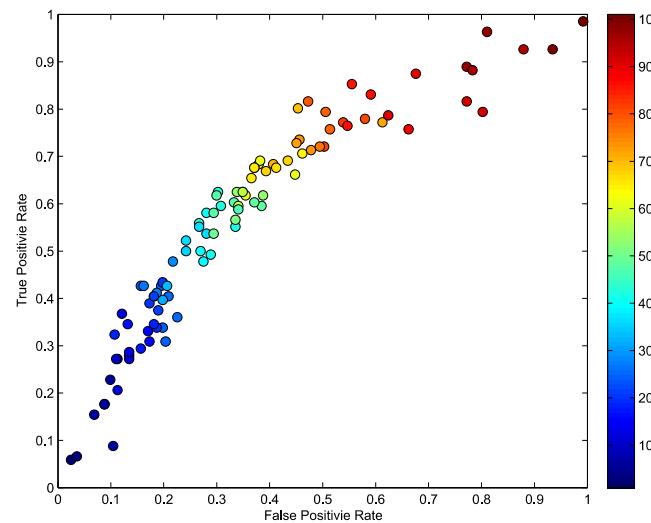
Intuición detrás del método



- ▶ En el método propuesto, los clasificadores base se fuerzan a aprender conjuntos de datos con mayor variabilidad en el ratio entre clases.
- ▶ Esto hace que los clasificadores base del ensemble se esparzan por el espacio ROC definido por FPR y TPR .
- ▶ Por lo tanto, se espera que sean más diversos y mejoren el rendimiento del ensemble.



Bagging



Ensemble de Random Balance

Clasificadores base en el espacio ROC (credit-g dataset).

Algoritmos de construcción de ensambles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

Ensembles para problemas desequilibrados

56 Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras



Algoritmos de
construcción de
ensembles y conjuntos
desequilibrados

José Francisco Díez
Pastor

Topics

Introducción a la
minería de datos

Ensembles para
problemas
desequilibrados

57 Contribuciones al
aprendizaje de
problemas
desequilibrados

Aplicaciones

Resumen y líneas
futuras

Existen varias modificaciones de AdaBoost.M2 para conjuntos desequilibrados. Las más conocidas son SMOTEBoost y RUSBoost.

RB-Boost es una modificación de AdaBoost.M2, en cada iteración el dataset es modificado usando Random Balance.

Configuración experimental



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

Ensembles para problemas desequilibrados

58 Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

- ▶ Dos colecciones de conjuntos de datos fueron usadas: La colección HDDT y la colección KEEL.
- ▶ El árbol J48 (C4.4 configuración C4.4) fué elegido como clasificador base de todos los ensembles.
- ▶ Se usaron rankings promedio para la comparación entre multiples algoritmos.
- ▶ Debido al alto número de métodos probados, las comparaciones se dividieron por familias.
 - ▶ Cada familia contiene los ensembles dependiendo de la estrategia principal de generación de diversidad.
 - ▶ Distinguimos tres familias: Data-preprocessing-only, Bagging and Boosting.

Resultados



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

Ensembles para problemas desequilibrados

59 Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

Table : Número de métodos en cada familia. Posición de Random Balance en el ranking promedio. En parentesis el número de métodos significativamente equivalentes con el primero del ranking

	Data-processing family (11)	Bagging family (13)	Boosting family (9)
AUC	1 (1)	1 (1)	1 (2)
F-Measure	1 (4)	2	1 (1)
G-Mean	1 (1)	1 (1)	3

Resultados



Algoritmos de
construcción de
ensembles y conjuntos
desequilibrados

José Francisco Díez
Pastor

Topics

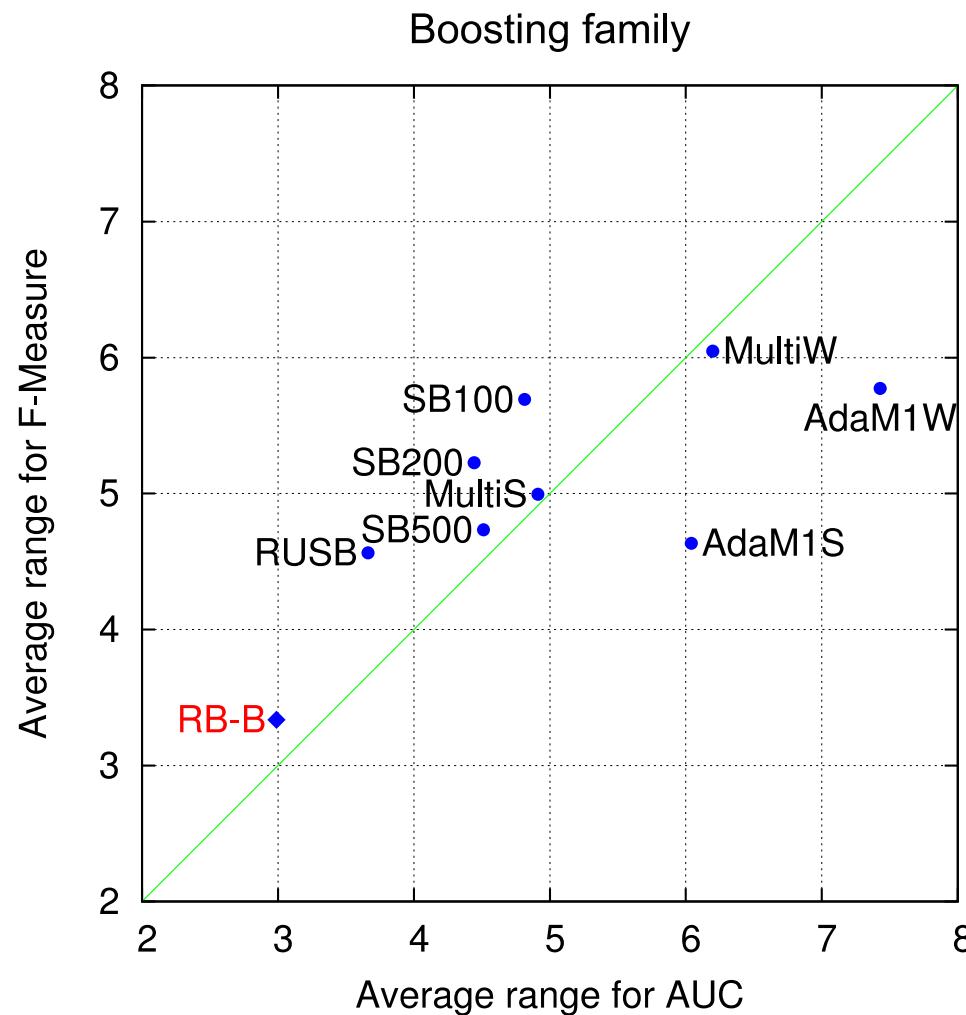
Introducción a la
minería de datos

Ensembles para
problemas
desequilibrados

60
Contribuciones al
aprendizaje de
problemas
desequilibrados

Aplicaciones

Resumen y líneas
futuras



Resultados



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

Ensembles para problemas desequilibrados

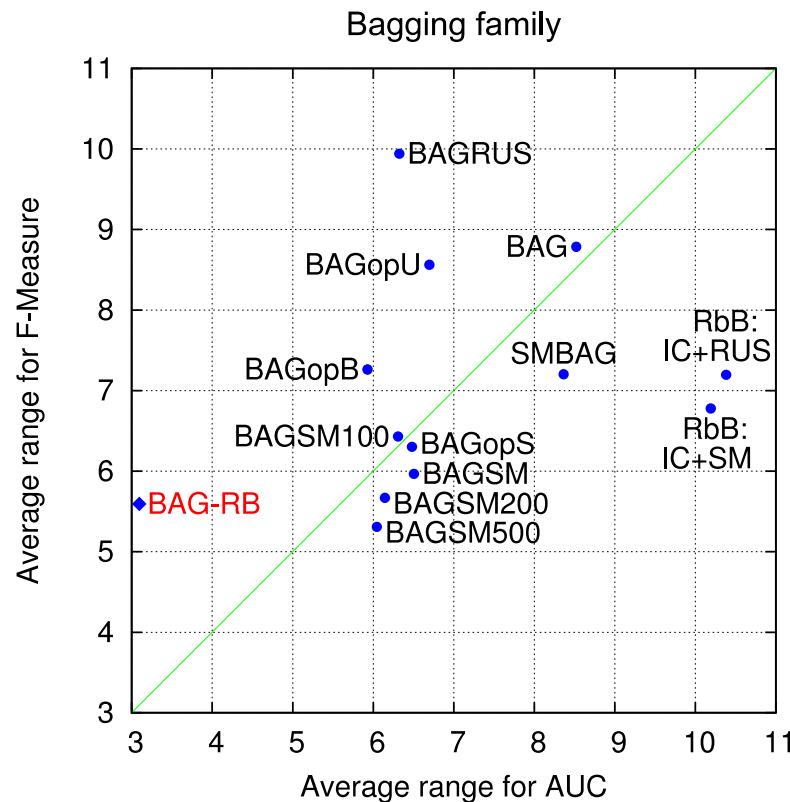
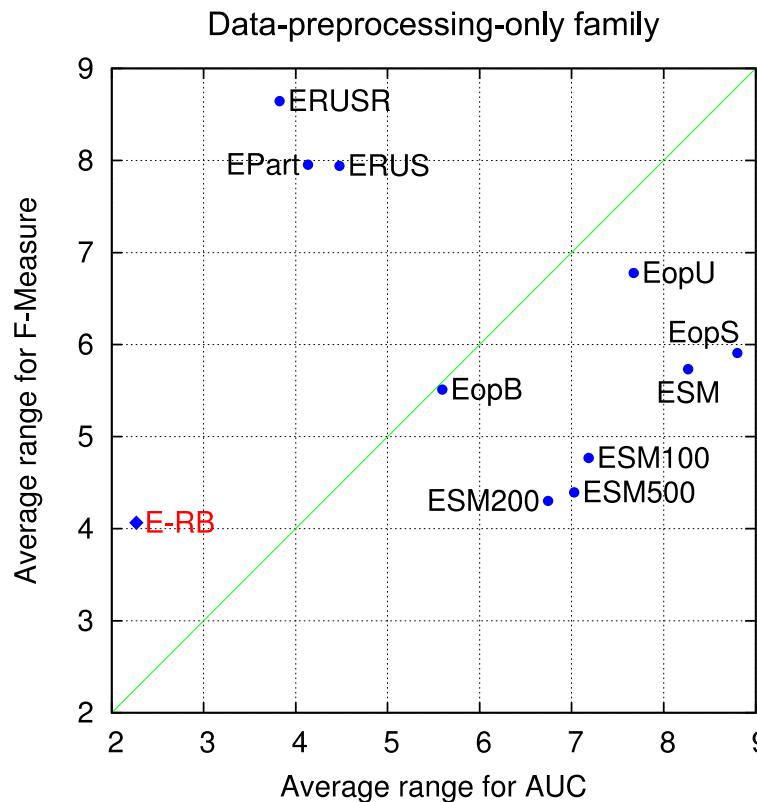
Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

61

Dept. Ingeniería civil
Universidad de Burgos



86

Random Balance: Conclusiones



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

Ensembles para problemas desequilibrados

62 Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

- ▶ Se ha propuesto una nueva técnica de preprocessado para conjuntos desequilibrados: Random Balance, basada en la idea de confiar en la aleatoriedad y la repetición para varying randomly the proportions of the classes, evitando la necesidad de ajustar la técnica y el ratio de oversampling o undersampling.
- ▶ Usando esta técnica creamos un nuevo ensemble: RB-Boost.
- ▶ A pesar de su simplicidad los métodos propuestos superan otros ensembles del estado del arte.

Diversity techniques for imbalance learning: Motivación



Algoritmos de
construcción de
ensembles y conjuntos
desequilibrados

José Francisco Díez
Pastor

Topics

Introducción a la
minería de datos

Ensembles para
problemas
desequilibrados

63 Contribuciones al
aprendizaje de
problemas
desequilibrados

Aplicaciones

Resumen y líneas
futuras

La mayoría de las técnicas de ensembles para conjuntos desequilibrados usan reweighting, oversampling o undersampling para reducir el ratio de desbalanceo.

Sin embargo Random Balance produce conjuntos de datos con un ratio aleatorio y los resultados son muy competitivos.

Diversity techniques for imbalance learning: Motivación



¿Cuál sería la influencia de las técnicas diseñadas para incrementar la diversidad en aprendizaje de conjuntos desequilibrados?

¿Funcionaría la combinación de técnicas de incremento de la diversidad con reweighing, oversampling o undersampling?

- ▶ **Díez-Pastor, J. F., Rodríguez, J. J., García-Osorio, C. & Kuncheva, L. (2015). Diversity techniques improve the performance of the best imbalance learning ensembles. Information Sciences, Accepted.**

Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

Ensembles para problemas desequilibrados

64 Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

Técnicas de diversidad



Algoritmos de construcción de ensambles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

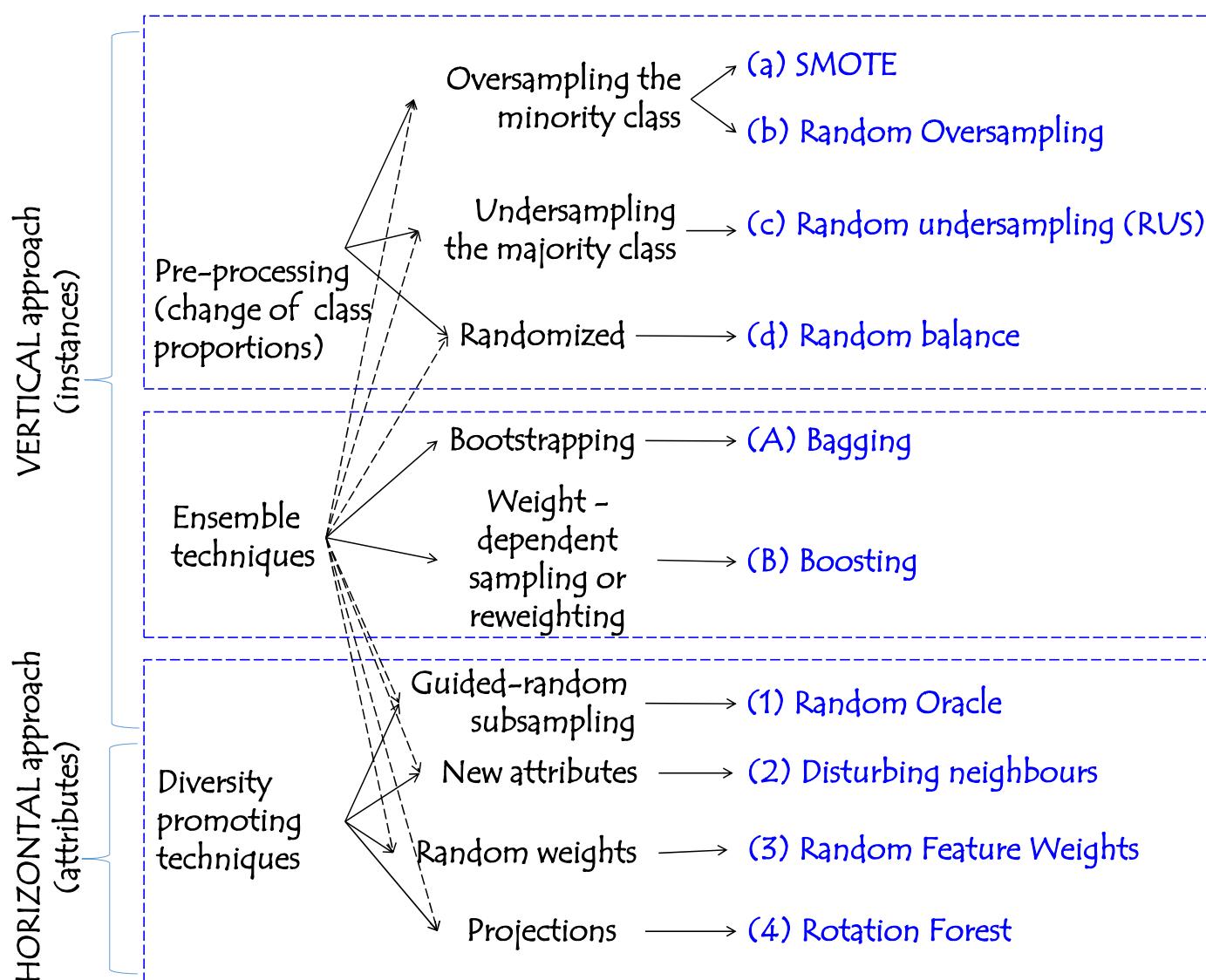
Introducción a la minería de datos

Ensembles para problemas desequilibrados

65 Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

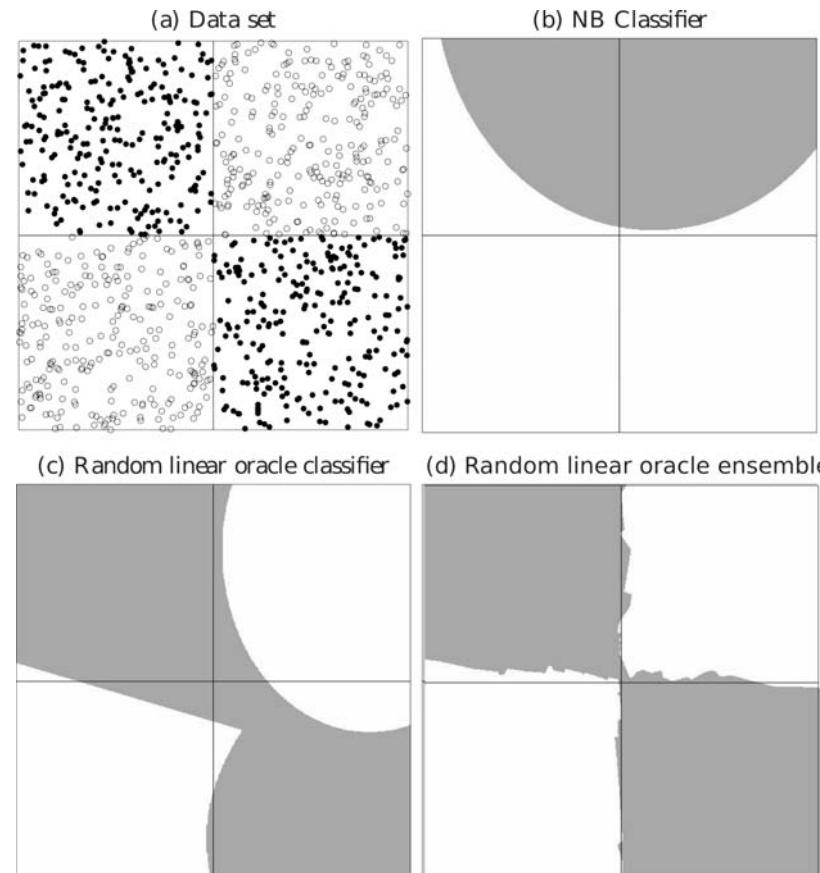
Resumen y líneas futuras



Técnicas de diversidad



Random Oracles En cada iteración (para cada clasificador base), el conjunto de datos es dividido en dos grupos usando un hyperplano aleatorio. Un clasificador diferente es construido con cada grupo.



Algoritmos de construcción de ensambles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

Ensembles para problemas desequilibrados

66 Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

Técnicas de diversidad



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

Ensembles para problemas desequilibrados

67 Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

Disturbing Neighbours En cada iteración N instancias son seleccionadas aleatoriamente (*Disturbing Neighbours*), a partir de estas, N atributos binarios son creados para cada instancia (con valor 1 si la correspondiente *Disturbing Neighbour* es el más cercano, 0 en caso contrario).

Técnicas de diversidad



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez
Pastor

Topics

Introducción a la minería de datos

Ensembles para problemas desequilibrados

68 Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

Random Feature Weights asocia un vector de pesos aleatorios a cada uno de los árboles del ensemble. Ese vector se usa para modificar el modo en el que se seleccionan los atributos en la construcción del árbol. Los atributos con un peso más alto son más probables que sean elegidos.

Rotation Forest usa análisis de componentes principales (PCA) para proyectar diferentes grupos de atributos. Los grupos de atributos y el subconjunto de instancias usado para calcular PCA es diferente en cada iteración.

Resultados experimentales



Se realizó un estudio experimental que combinaba estas técnicas especialmente diseñadas para incrementar la diversidad con aquellas técnicas diseñadas para trabajar con conjuntos desequilibrados.

- ▶ 84 conjuntos de datos de los repositorios de Keel y HDDT.
- ▶ 17 métodos del estado del arte para conjuntos desequilibrados: RAMOBoost, Random Balance Boost, RUSBoost, SMOTEBoost, etc.
- ▶ Las 4 técnicas de incremento de la diversidad anteriores.

Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

Ensembles para problemas desequilibrados

69 Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

Resultados experimentales



Algoritmos de construcción de ensambles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

Ensembles para problemas desequilibrados

70 Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

Abbreviation	Name	Pre-processing				Ensemble		Type		
		a	b	c	d	A	B	x	y	z
E-SM100	Ensemble SMOTE 100%	■						■		
E-SM	Ensemble SMOTE	■						■		
E-RUS	Ensemble RUS			■				■		
E-RB	Ensemble Random Balance	■			■	■			■	
Ba	Bagging					■		■		
SMBa	SMOTEBagging	■	■			■			■	
Ba-SM100	Bagging SMOTE 100%	■				■			■	
Ba-SM	Bagging SMOTE	■				■			■	
Ba-RUS	Bagging RUS			■		■			■	
Ba-RB	Bagging Random Balance	■		■	■	■			■	
ABo1	AdaBoost. M1						■	■		
ABo2	AdaBoost.M2						■	■		
MBo	MultiBoost						■		■	
SMBo	SMOTEBoost	■					■		■	
RAMOBo	RAMOBoost	■					■		■	
RUSBo	RUSBoost			■		■			■	
RBBo	Random Balance-Boost	■		■	■	■			■	

a) SMOTE, b) Random Oversampling, c) Random Undersampling, d) Random Balance. A) Bagging, B) Boosting.

Type: x) baseline methods which are not designed to deal with imbalanced data, y) ensemble methods especially designed to deal with imbalanced data, and z) baseline methods combined with preprocessing techniques.

Resultados experimentales



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

Ensembles para problemas desequilibrados

71 Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

Varios diferentes análisis fueron realizados en este artículo, de los que se mencionarán tres.

1. Efectos de las técnicas de diversidad en combinación con ensembles para desequilibrados
2. Determinación de cual es la mejor técnica de acuerdo a las distintas métricas.
3. Uso de medidas de complejidad para predecir cuando es más conveniente usar técnicas de diversidad para mejorar los resultados.

Técnicas de diversidad en combinación con ensembles para desequilibrados



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

Ensembles para problemas desequilibrados

72 Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

AUC

	Normal	(1) Oracles	(2) Disturbing Neighbours	(3) RFW	(4) Rotation Forest
Ensemble	2.554	2.679	3.381	[1.387]	
E-SM100	3.298	2.810	3.107	3.476	[2.310]
E-SM	4.185	[1.929]	2.887	3.643	(2.357)
E-RUS	4.494	(2.720)	(2.357)	3.143	[2.286]
E-RB	4.571	2.631	3.048	2.875	[1.875]
Ba	4.548	2.720	2.631	3.280	[1.821]
SMBa	4.702	2.720	3.327	2.601	[1.649]
Ba-SM100	4.524	2.863	2.732	2.827	[2.054]
Ba-SM	4.506	2.679	3.089	2.732	[1.994]
Ba-RUS	4.179	(2.702)	(2.512)	3.143	[2.464]
Ba-RB	4.274	(2.685)	2.940	2.887	[2.214]
ABo1	3.518	(2.869)	(2.887)	(3.131)	[2.595]
ABo2	4.417	2.774	3.131	2.923	[1.756]
MBo	3.744	(2.726)	(2.804)	3.262	[2.464]
SMBo	4.286	2.804	3.083	2.857	[1.970]
RAMOBo	4.339	2.940	2.976	2.845	[1.899]
RUSBo	3.929	2.940	[2.208]	3.327	(2.595)
RBBo	3.732	3.095	(2.786)	3.173	[2.214]

Average ranks

Técnicas de diversidad en combinación con ensembles para desequilibrados



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

73

	Normal	(1) Oracles	(2) Disturbing Neighbours	(3) RFW	(4) Rotation Forest
Ensemble		(2.411) (2.542) (2.738)	[2.310]		
E-SM100	(2.833) (2.810) [2.762]	3.661	(2.935)		
E-SM	3.476 [2.232]	2.827	3.607	2.857	
E-RUS	4.298	2.845	2.643	3.077	[2.137]
E-RB	3.923	[2.619] (2.679)	(3.060)	(2.720)	
Ba	(2.964) (2.756) [2.500]	3.530	3.250		
SMBa	4.167	(2.589)	3.202	2.917	[2.125]
Ba-SM100	3.405	(2.714) [2.702]	3.351	(2.827)	
Ba-SM	3.815	(2.536) (2.845)	3.357	[2.446]	
Ba-RUS	4.244	3.006	(2.536)	2.988	[2.226]
Ba-RB	3.905	(2.685) [2.530]	(3.113)	(2.768)	
ABo1	(2.899) (2.798) [2.518]	3.756	(3.030)		
ABo2	3.833	[2.476] (2.607)	3.542	(2.542)	
MBo	(2.851) (2.786) [2.565]	3.524	3.274		
SMBo	3.917	(2.548) (2.696)	3.470	[2.369]	
RAMOBo	3.869	2.744	(2.577)	3.619	[2.190]
RUSBo	3.815	2.958	(2.435)	3.595	[2.196]
RBBo	3.815	(2.685) (2.530)	3.595	[2.375]	

Average ranks

	Normal	(1) Oracles	(2) Disturbing Neighbours	(3) RFW	(4) Rotation Forest
Ensemble		[2.280] (2.304)	(2.536)	2.881	
E-SM100	(2.762) (2.762) [2.690]	3.708	3.077		
E-SM	3.298	[2.387]	2.875	3.988	(2.452)
E-RUS	4.369	(2.595) (2.643)	3.137	[2.256]	
E-RB	3.577	[2.476] (2.702)	3.679	(2.565)	
Ba	(2.488)	2.958	[2.321]	3.554	3.679
SMBa	3.619	[2.446] (2.940)	3.500	(2.494)	
Ba-SM100	(2.833) (2.857) [2.488]	3.482	3.339		
Ba-SM	3.363	[2.298] (2.798)	4.012	(2.530)	
Ba-RUS	3.982	[2.649] (2.690)	(2.762)	(2.917)	
Ba-RB	3.500	[2.506] (2.613)	3.685	(2.696)	
ABo1	(2.708) (2.714) [2.649]	3.768	(3.161)		
ABo2	3.262	(2.476) [2.405]	3.673	3.185	
MBo	(2.649) (2.786) [2.565]	3.595	3.405		
SMBo	3.536	(2.560) [2.446]	3.625	(2.833)	
RAMOBo	3.393	(2.744) [2.542]	3.655	(2.667)	
RUSBo	3.054	3.101	[2.185]	3.107	3.554
RBBo	3.708	(2.637) [2.363]	3.607	(2.685)	

Average ranks

F-Measure

G-Mean

Determinación de cual es la mejor técnica de acuerdo a las distintas métricas.



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

Ensembles para problemas desequilibrados

74 Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

AUC El ganador global es Rotation Forest.

F-Measure El mejor método es RAMOBoost combinado con Rotation Forest.

G-Mean El mejor método es Bagging+RUS combinado con Random Oracles.

Las mejores combinaciones de acuerdo a la F-measure usan oversampling mientras que las mejores de acuerdo a la G-mean usan undersampling.

El mejor método de acuerdo a la AUC no usa ninguna técnica de preprocessado para desequilibrados.

Cuando es conveniente usar técnicas de diversidad usando medidas de complejidad?



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez
Pastor

Topics

Introducción a la minería de datos

Ensembles para problemas desequilibrados

75 Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

Resumen y líneas futuras

El conjunto de meta-características fue creado de la siguiente manera:

- ▶ 5712 instancias: $84 \text{ datasets} \times 17 \text{ ensembles} \times 4 \text{ técnicas de diversidad.}$
- ▶ Cada instancia:
 - ▶ 14 medidas de complejidad extraídas del dataset (usando DCol), nombre de ensemble m , nombre de la técnica de diversidad d .
 - ▶ clase: (True/False) si $m + d$ mejora m en el dataset.

HotSpot fue usado para encontrar un conjunto de reglas entre las medidas de complejidad y la clase.

Cuando es conveniente usar técnicas de diversidad usando medidas de complejidad?



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

Ensembles para problemas desequilibrados

76 Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

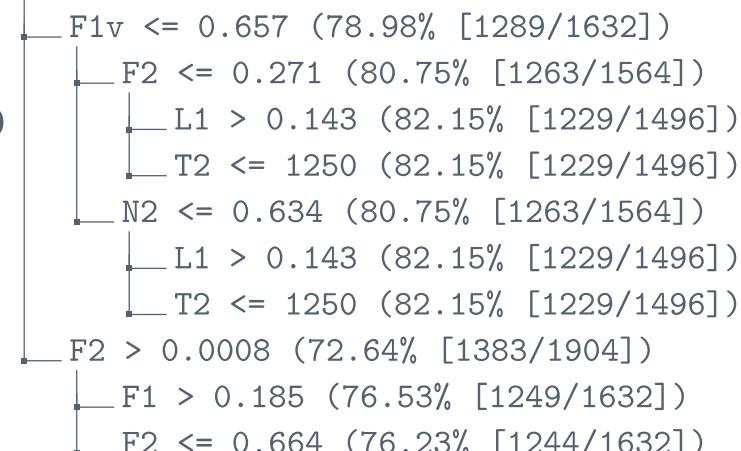
Resumen y líneas futuras

Se encontraron algunas reglas interesantes, por ejemplo:

- ▶ Si el solape entre *per-class bounding boxes* (F_2) es alto o *maximum Fisher's discriminant ratio* (F_1) es bajo, es una buena idea aplicar técnicas de incremento de la diversidad.

Reglas para la F-Measure

Class=yes (65% [3713/5712])



Diversity techniques for imbalance learning: Conclusiones



Algoritmos de
construcción de
ensembles y conjuntos
desequilibrados

José Francisco Díez
Pastor

Topics

Introducción a la
minería de datos

Ensembles para
problemas
desequilibrados

77 Contribuciones al
aprendizaje de
problemas
desequilibrados

Aplicaciones

Resumen y líneas
futuras

- ▶ Los ensembles que usan técnicas de incremento de la diversidad obtienen mejor ranking que su contraparte que no la usa. Curioso porque ninguna técnicas afecta al desbalanceo.
- ▶ Los resultados obtenidos para una medida de rendimiento no pueden ser extrapolados a otras.



Algoritmos de
construcción de
ensembles y conjuntos
desequilibrados

José Francisco Díez
Pastor

Topics

Introducción a la
minería de datos

Ensembles para
problemas
desequilibrados

Contribuciones al
aprendizaje de
problemas
desequilibrados

78 Aplicaciones

Resumen y líneas
futuras

Aplicaciones

Applications



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

79 Aplicaciones

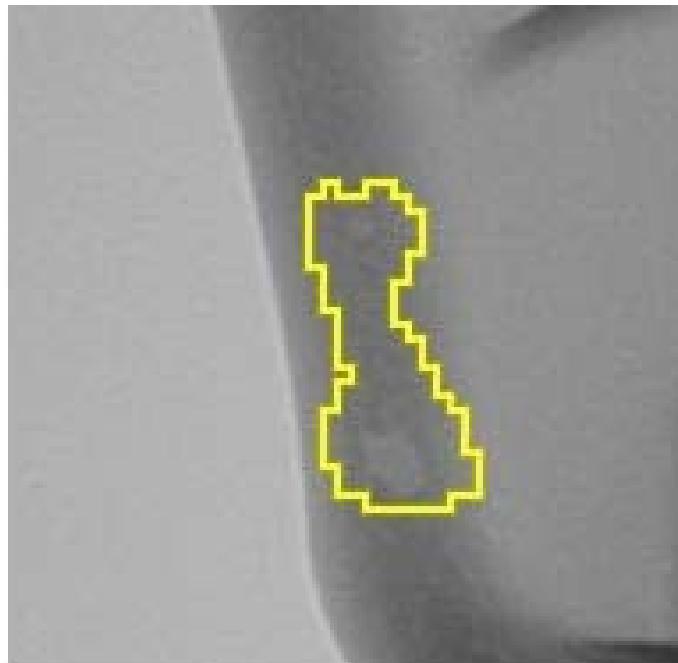
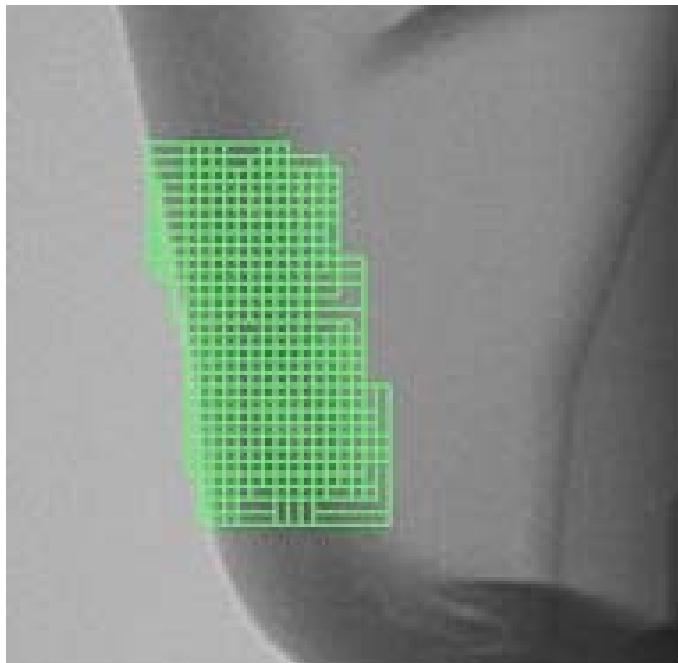
Resumen y líneas futuras

- ▶ ¿Es posible usar ensembles para conjuntos desequilibrados en la detección de defectos de fabricación?
- ▶ Díez-Pastor, J. F., García-Osorio, C., Barbero-García, V., & Blanco-Álamo, A. (2013). Imbalanced Learning Ensembles for Defect Detection in X-Ray Images. In Recent Trends in Applied Artificial Intelligence (pp. 654-663). Springer Berlin Heidelberg.
- ▶ Díez-Pastor, J. F., Arnaiz-González, A., García-Osorio, C., & Rodríguez-Diez, J. J. Segmentación de defectos en piezas de fundido usando umbrales adaptativos y ensembles (Segmentation of defects in castings using adaptive thresholds and ensembles). XVII Congreso Español sobre tecnologías y lógica fuzzy (ESTYLF 2014) pp. 345-349.

Detección de defectos en imágenes de rayos X



- ▶ Realizamos la detección de defectos en piezas metálicas analizando imágenes de rayos X con dos estrategias:
 - ▶ Ventana deslizante: Cada ventana es una instancia, cada pixel está cubierto por varias ventanas, usando un umbral se puede decidir cuando un pixel pertenece a un defecto o no.



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez
Pastor

Topics

Introducción a la minería de datos

Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

80

Aplicaciones

Resumen y líneas futuras

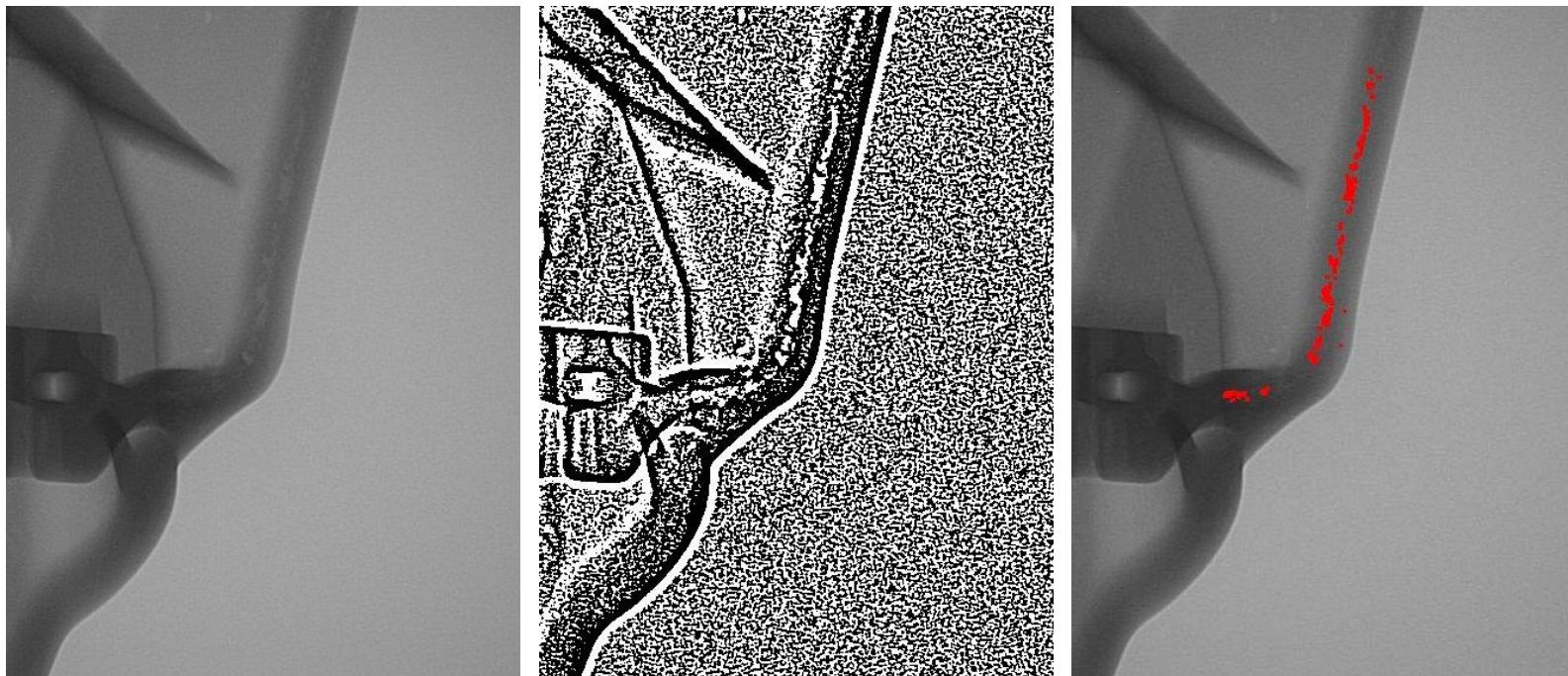
86

Detección de defectos en imágenes de rayos X



► Segunda estrategia:

- Un procedimiento en dos pasos: determinar las regiones susceptibles de ser defecto y posteriormente clasificarlas como defecto o no.



Algoritmos de construcción de ensambles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

81

Aplicaciones

Resumen y líneas futuras

86



Algoritmos de
construcción de
ensembles y conjuntos
desequilibrados

José Francisco Díez
Pastor

Topics

Introducción a la
minería de datos

Ensembles para
problemas
desequilibrados

Contribuciones al
aprendizaje de
problemas
desequilibrados

Aplicaciones

82 Resumen y líneas
futuras

Resumen y líneas futuras

Resumen



Algoritmos de construcción de ensembles y conjuntos desequilibrados

José Francisco Díez Pastor

Topics

Introducción a la minería de datos

Ensembles para problemas desequilibrados

Contribuciones al aprendizaje de problemas desequilibrados

Aplicaciones

83 Resumen y líneas futuras

- ▶ Basado en la idea de variar aleatoriamente las proporciones entre clases se ha creado un ensemble para conjuntos desequilibrados que supera el otros métodos del estado del arte.

Resumen



Algoritmos de
construcción de
ensembles y conjuntos
desequilibrados

José Francisco Díez
Pastor

Topics

Introducción a la
minería de datos

Ensembles para
problemas
desequilibrados

Contribuciones al
aprendizaje de
problemas
desequilibrados

Aplicaciones

Resumen y líneas
futuras

84

- ▶ Se ha comprobado experimentalmente que las técnicas de incremento de la diversidad mejoran los ensembles específicos para conjuntos desequilibrados.
- ▶ Se han usado ensembles para resolver un problema real en la industria manufacturera.

86

Líneas futuras



Algoritmos de
construcción de
ensembles y conjuntos
desequilibrados

José Francisco Díez
Pastor

Topics

Introducción a la
minería de datos

Ensembles para
problemas
desequilibrados

Contribuciones al
aprendizaje de
problemas
desequilibrados

Aplicaciones

85 Resumen y líneas
futuras

Hay varias líneas futuras

- ▶ Estudiar el rendimiento de estrategias aleatorias en presencia de problemas típicos de conjuntos desequilibrados como el ruido, solape, *small disjuncts*, ejemplos *borderline* etc. Existen técnicas de preprocessado apropiadas para estos problemas, pero su elección y parámetros es complicada.
- ▶ Encontrar la combinación óptima de preprocessado y estrategia de diversidad para cada conjunto de datos usando ciertas meta-características.



Algoritmos de
construcción de
ensembles y conjuntos
desequilibrados

José Francisco Díez
Pastor

Topics

Introducción a la
minería de datos

Ensembles para
problemas
desequilibrados

Contribuciones al
aprendizaje de
problemas
desequilibrados

Aplicaciones

86 Resumen y líneas
futuras

Preguntas

Gracias por su atención
Plantilla de Beamer realizada por Jesper Kjær Nielsen

