UNIVERSIDAD DE BURGOS
ESCUELA DE DOCTORADO

TESIS DOCTORALES

TÍTULO: ESTUDIO DE MÉTODOS DE CONSTRUCCIÓN DE ENSEMBLES DE CLASIFICADORES Y APLICACIONES

AUTOR: DÍEZ PASTOR, JOSÉ FRANCISCO
PROGRAMA DE DOCTORADO: INVESTIGACIÓN EN INGENIERÍA
FECHA LECTURA: 15/06/2015
HORA: 12:00
CENTRO LECTURA: ESCUELA POLITÉCNICAS SUPERIOR. SALA DE JUNTAS 2. CAMPUS DEL VENA
DIRECTORES: CÉSAR IGNACIO GARCÍA OSORIO y JUAN JOSÉ RODRÍGUEZ DÍEZ
TRIBUNAL: COLIN FYFE
JESÚS MANUEL MAUDES RAEDO
NICOLÁS GARCÍA PEDRAJAS
MACIEJ STANISLAW GRZENDA
JULIÁN LUENGO MARTÍN

RESUMEN:

La inteligencia artificial es el área de conocimiento que se dedica a la creación de sistemas informáticos con un comportamiento inteligente. Dentro de este área se puede considerar que el aprendizaje computacional estudia la creación de sistemas que aprenden por sí mismos. En el aprendizaje supervisado se le proporcionan al sistema tanto las entradas como la salida esperada, cuando la salida es de tipo categórico, se trata de un clasificador y cuando la salida es numérica, se trata de un regresor. En ocasiones en ciertos problemas ocurre que el número de ejemplos de un tipo es mucho mayor que el número de ejemplos de otro tipo, cuando esto ocurre se habla de conjuntos desequilibrados. La combinación de varios clasificadores o regresores es lo que se denomina ensemble, y a menudo ofrece mejores resultados que cualquiera de los miembros que lo forman.

Esta tesis, se centra en el desarrollo de nuevos algoritmos de construcción de ensembles, sobre todo haciendo hincapié a las técnicas de incremento de la diversidad en ensembles homogéneos (cuando todos los miembros están construidos usando la misma técnica).

En la primera parte de la tesis, se presenta un breve estudio de los métodos más representativos de las distintas técnicas de construcción de ensembles, aprendizaje en conjuntos desequilibrados y breves nociones sobre validación experimental. En la segunda, aparecen todas las publicaciones que han sido realizadas en el contexto de esta tesis. Esta segunda parte puede ser dividida en tres bloques.

En un primer bloque, se explora la utilización de la fase constructiva de la metaheurística GRASP como una manera de inyectar aleatoriedad en algoritmos de construcción de árboles. Inyectar aleatoriedad en el propio algoritmo del clasificador base es una de las técnicas usadas para incrementar la diversidad de un ensemble. Esta técnica, que ha sido llamada "GRASP Forest" ha sido utilizada con éxito en árboles de clasificación y árboles de regresión.

La diversidad es clave en los ensembles, pero se quiere incrementar la diversidad sin afectar gravemente a la precisión de los clasificadores base. Profundizando en la idea anterior, se ha desarrollado un segundo método "GAR-Forest" GRASP with annealed randomness.

En este método se parte de la idea intuitiva de que los nodos que más influencia tienen en la correcta clasificación de las instancias son los nodos inferiores y hojas, mientras que los nodos que más afectan la estructura global del árbol (y por lo tanto la diversidad) son la raíz y los nodos superiores. Partiendo de esa idea se ha diseñado un método que utiliza la metaheurística GRASP, para controlar el nivel de aleatoriedad en cada uno de los niveles en el árbol. Creando árboles en donde la raíz es completamente aleatoria y el nivel de aleatoriedad que disminuye según se construye el árbol.

En un segundo bloque se aborda el problema de los ensembles para conjuntos desequilibrados. Existen varias estrategias para lidiar con el problema del desequilibrio, una de ellas es utilizar
distintos métodos de preprocesado como Undersampling o SMOTE, los parámetros óptimos de estos métodos son dependientes del problema y a menudo son difíciles de encontrar. En este bloque se presenta un método llamado "Random Balance", basado en la idea de variar aleatoriamente las proporciones entre las clases,

confiando en esta heurística, se elimina la necesidad de ajustar parámetros, a la vez que se aumenta la diversidad del ensemble.

Como se ha mencionado previamente, cuando se aborda el problema del desequilibrio, las técnicas más comúnmente usadas son aquellas que afectan la proporción entre las clases (re-pesado, sobremuestreo y submuestreo, etc.).
En otro trabajo de este bloque se estudia el efecto de distintas técnicas (Random Oracles, Random Feature Weights, Rotation Forest y Disturbing
Neighbours), originalmente destinadas a aumentar la diversidad en ensembles no desequilibrados. Se realizan varios análisis sobre el impacto del tamaño del ensemble en el rendimiento del ensemble, se estudia en qué ocasiones estas técnicas mejoran a los ensembles del estado del arte para desequilibrados y qué combinación de ensemble y técnica de diversidad es la que ofrece mejores resultados para distintas medidas. También se realiza un estudio que trata de predecir en qué ocasiones es más apropiado utilizar técnicas de incremento de la diversidad basándose en distintas meta-características propias del conjunto de datos.

Finalmente, se aplican algunas de estas técnicas a la solución de varias aplicaciones reales. Se han aplicado ensembles para la predicción de la calidad superficial en procesos de mecanizado y para el desarrollo de un sistema de detección de defectos en piezas metálicas mediante imágenes de radiografía.

PALABRAS CLAVE:
Minería de datos, ensembles, diversidad, GRASP, Random Balance, Boosting projections, boosting, conjuntos de datos desequilibrados, análisis de radiografía