Métodos de Conjunto en Aprendizaje Automático

Pedro Javier Loor Delgado

17 de octubre de 2023

Resumen

Los métodos de conjunto son algoritmos que combinan varios clasificadores para predecir nuevas clasificaciones. Este artículo revisa estos métodos y por que funcionan mejor que un solo clasificador. También se analizan investigaciones previas y se presentan experimentos sobre Adaboost y el sobreajuste.

1. Introducción

En el aprendizaje supervisado, los métodos de conjunto combinan múltiples clasificadores para predecir nuevas clasificaciones. Esto se basa en el principio de que los conjuntos suelen ser mas precisos que los clasificadores individuales. Para que un conjunto sea efectivo, los clasificadores deben ser precisos y diversificados, es decir, cometer errores diferentes en datos nuevos. La combinación de sus decisiones, a menudo mediante votación ponderada, mejora la precision general. La diversidad y precisión son cruciales para el éxito de los conjuntos.

El éxito de los métodos de conjunto radica en la combinación de clasificadores individuales con tasas de error bajos y no correlacionados. Esto reduce el riesgo de elegir un clasificador incorrecto. Además, la falta de datos de entrenamiento en comparación con el espacio de hipotesis hace que la construcción de conjuntos sea efectiva, ya que varios clasificadores precisos pueden dar resultados similares y promediar sus votos mejora la precisión.

Existen tres razones fundamentales por las cuales los metodos de conjunto son efectivos en el aprendizaje automatico.

Primero, los conjuntos pueden mejorar la precision cuando las hipotesis individuales tienen tasas de error bajas y no correlacionadas.

Segundo, en problemas computacionalmente dificiles, los conjuntos pueden proporcionar aproximaciones mas efectivas al combinar multiples busquedas locales.

Tercero, los conjuntos pueden ampliar el espacio de funciones representables al combinar hipotesis de manera ponderada, lo que resulta en una mayor flexibilidad en la representacion de funciones. Los metodos de conjunto tienen el potencial de superar deficiencias comunes en los algoritmos de aprendizaje tradicionales.

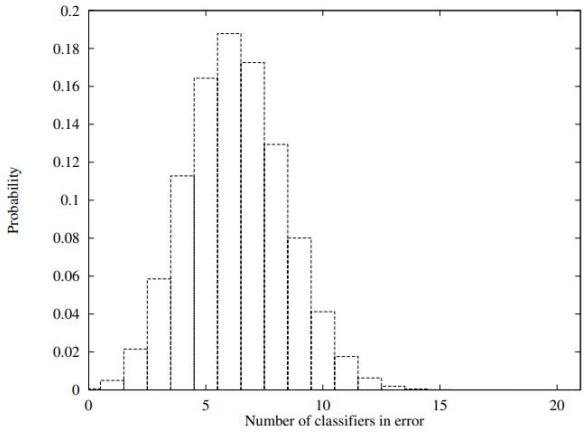

Figura 1: La probabilidad de que exactamente f (de 21) hipótesis cometan un error, suponiendo que cada hipótesis tiene una tasa de error de 0.3 y comete sus errores de manera independiente de las otras hipótesis.

2. Métodos para Construir Conjuntos

Revisaremos metodos de construccion de conjuntos que son versatiles y aplicables a diversos algoritmos de aprendizaje.

2.1. Votación Bayesiana: Enumerando las Hipótesis

En un contexto bayesiano, las hipótesis definen distribuciones de probabilidad condicional. Se aborda el problema de predicción como una combinación ponderada de estas hipótesis, utilizando probabilidades a priori y verosimilitudes. En entornos con pocos datos de entrenamiento, el enfoque bayesiano puede mejorar la precisión al promediar las hipótesis. Sin embargo, en problemas complejos, donde no se pueden enumerar todas las hipótesis, se utilizan aproximaciones como el muestreo aleatorio. La elección de la representación y la probabilidad a priori es fundamental, y en algunos casos, otros métodos de conjunto pueden ser más efectivos. El enfoque bayesiano no aborda problemas computacionales y de representación de manera significativa.

2.2. Manipulando los Ejemplos de Entrenamiento

El segundo método para construir conjuntos implica manipular ejemplos de entrenamiento para generar múltiples hipótesis. Esto es especialmente útil para algoritmos de aprendizaje inestables. Bagging es una técnica común que utiliza muestras de entrenamiento aleatorias con reemplazo. Otra técnica deja fuera subconjuntos disjuntos de datos. El tercer método, ilustrado por AdaBoost, ajusta los pesos de los ejemplos de entrenamiento en cada iteracion para

construir problemas de aprendizaje mas difíciles. AdaBoost es un algoritmo por etapas que busca minimizar un error específico o maximizar el margen en los datos de entrenamiento.

2.3. Manipulación de las Características de Entrada

Una técnica general para generar múltiples clasificadores es manipular las características de entrada disponibles para el algoritmo de aprendizaje. Esto se hace creando diferentes conjuntos de características y tamaños de red neuronal, por ejemplo. Funciona mejor cuando las características de entrada son redundantes.

2.4. Manipulación de los Objetivos de Salida

Una técnica general para construir conjuntos de clasificadores implica manipular los valores ƨ"que se proporcionan al algoritmo de aprendizaje. Esto se logra dividiendo las clases en dos subconjuntos y relabelando los datos para generar múltiples clasificadores. Cada clasificador vota por una clase y se elige la clase con más votos como la predicción del conjunto. Este enfoque se ha utilizado con éxito en problemas de clasificación difíciles y puede funcionar con cualquier algoritmo de aprendizaje para problemas de 2 clases. También se ha combinado con técnicas de selección de características para mejorar el rendimiento en tareas específicas.

2.5. Inyectando Aleatoriedad

Inyectar aleatoriedad en el algoritmo de aprendizaje es una técnica efectiva para generar conjuntos de clasificadores. Se puede aplicar a algoritmos de redes neuronales, árboles de decisión y otros. Esta aleatoriedad se introduce a través de la inicialización de pesos, selección aleatoria de características o perturbación de datos de entrada. Estos conjuntos aleatorios a menudo superan el rendimiento de un solo clasificador y son útiles en una variedad de tareas de clasificación.

3. Comparando Diferentes Métodos de Conjuntos

Los estudios comparativos de métodos de conjuntos muestran que AdaBoost a menudo supera a otros métodos, como Bagging y árboles aleatorios, en condiciones de bajo ruido en los datos. Sin embargo, cuando se introduce ruido significativo, Bagging demuestra ser más robusto y efectivo que AdaBoost. La explicación radica en las diferencias en como abordan los problemas estadísticos, computacionales y de representación de los algoritmos base. AdaBoost busca optimizar el voto ponderado directamente, lo que lo hace propenso al sobreajuste en presencia de ruido, mientras que Bagging y árboles aleatorios se centran en abordar principalmente el problema estadístico y, por lo tanto, son más robustos

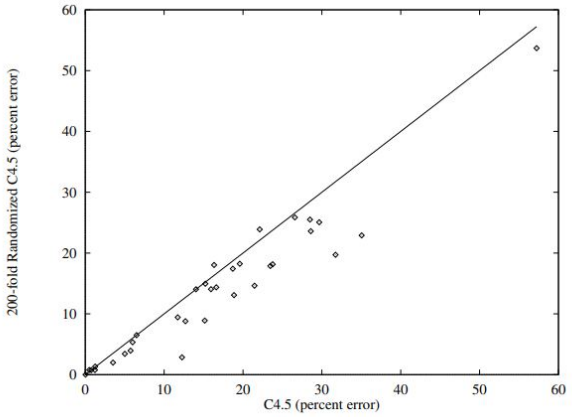

Figura 2: Comparación de la tasa de error de C4.5 con un conjunto de 200 árboles de decisión construidos mediante la introducción de aleatoriedad en C4.5 y luego realizando una votación uniforme.

frente al ruido. La etapa por etapa de AdaBoost tambien contribuye a su éxito en problemas de baja complejidad.

4. Conclusiones

Los conjuntos son una forma efectiva de mejorar la precisión de los clasificadores al combinar múltiples clasificadores menos precisos. Este artículo resume métodos para crear conjuntos y analiza por que funcionan bien, con un enfoque en el éxito de AdaBoost.

Un tema abierto sin explorar en este artículo es como AdaBoost interactúa con las características de los algoritmos de aprendizaje subyacentes. La mayoría de los algoritmos emparejados con AdaBoost son globales, y sería interesante investigar si algoritmos locales pueden combinar de manera efectiva con AdaBoost para crear nuevos y emocionantes enfoques de aprendizaje.

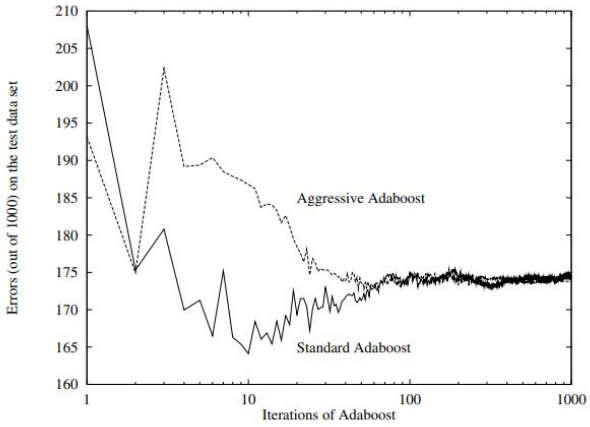

Figura 3: AdaBoost Agresivo muestra un rendimiento mucho peor que AdaBoost Estándar en un problema sintético desafiante.