

Metodos de Conjunto en Aprendizaje Automatico

Pedro Javier Loor Delgado

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

1 Introduccion

En el aprendizaje supervisado, los metodos de conjunto combinan multiples 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 combinacion de sus decisiones, a menudo mediante votacion ponderada, mejora la precision general. La diversidad y precision son cruciales para el exito de los conjuntos.

El exito de los metodos de conjunto radica en la combinacion de clasificadores individuales con tasas de error bajos y no correlacionados. Esto reduce el riesgo de elegir un clasificador incorrecto. Ademas, la falta de datos de entrenamiento en comparacion con el espacio de hipotesis hace que la construccion de conjuntos sea efectiva, ya que varios clasificadores precisos pueden dar resultados similares y promediar sus votos mejora la precision.

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.

2 Metodos para Construir Conjuntos

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

2.1 Votacion Bayesiana: Enumerando las Hipotesis

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

2.2 Manipulando los Ejemplos de Entrenamiento

El segundo metodo para construir conjuntos implica manipular ejemplos de entrenamiento para generar multiples hipotesis. Esto es especialmente util para algoritmos de aprendizaje inestables. Bagging es una tecnica comun que utiliza muestras de entrenamiento aleatorias con reemplazo. Otra tecnica deja fuera subconjuntos disjuntos de datos. El tercer metodo, ilustrado por AdaBoost, ajusta los pesos de los ejemplos de entrenamiento en cada iteracion para construir problemas de aprendizaje mas dificiles. AdaBoost es un algoritmo por etapas que busca minimizar un error especifico o maximizar el margen en los datos de entrenamiento.

2.3 Manipulacion de las Caracteristicas de Entrada

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

2.4 Manipulacion de los Objetivos de Salida

Una tecnica general para construir conjuntos de clasificadores implica manipular los valores "y" que se proporcionan al algoritmo de aprendizaje. Esto se logra dividiendo las clases en dos subconjuntos y relabelando los datos para generar multiples clasificadores. Cada clasificador vota por una clase y se elige la clase con mas votos como la prediccion del conjunto. Este enfoque se ha utilizado con exito en problemas de clasificacion dificiles y puede funcionar con cualquier algoritmo de aprendizaje para problemas de 2 clases. Tambien se ha combinado con tecnicas de seleccion de caracteristicas para mejorar el rendimiento en tareas especificas.

2.5 Inyectando Aleatoriedad

Inyectar aleatoriedad en el algoritmo de aprendizaje es una tecnica efectiva para generar conjuntos de clasificadores. Se puede aplicar a algoritmos de redes neuronales, arboles de decision y otros. Esta aleatoriedad se introduce a traves de la inicializacion de pesos, seleccion aleatoria de caracteristicas o perturbacion de datos de entrada. Estos conjuntos aleatorios a menudo superan el rendimiento de un solo clasificador y son utiles en una variedad de tareas de clasificacion.

3 Comparando Diferentes Metodos de Conjuntos

Los estudios comparativos de metodos de conjuntos muestran que AdaBoost a menudo supera a otros metodos, como Bagging y arboles aleatorios, en condiciones de bajo ruido en los datos. Sin embargo, cuando se introduce ruido significativo, Bagging demuestra ser mas robusto y efectivo que AdaBoost. La explicacion radica en las diferencias en como abordan los problemas estadisticos, computacionales y de representacion 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 arboles aleatorios se centran en abordar principalmente el problema estadistico y, por lo tanto, son mas robustos frente al ruido. La etapa por etapa de AdaBoost tambien contribuye a su exito en problemas de baja complejidad.

4 Conclusiones

Los conjuntos son una forma efectiva de mejorar la precision de los clasificadores al combinar multiples clasificadores menos precisos. Este articulo resume metodos para crear conjuntos y analiza por que funcionan bien, con un enfoque en el exito de AdaBoost.

Un tema abierto sin explorar en este articulo es como AdaBoost interactua con las caracteristicas de los algoritmos de aprendizaje subyacentes. La mayoria de los algoritmos emparejados con AdaBoost son globales, y seria interesante investigar si algoritmos locales pueden combinar de manera efectiva con AdaBoost para crear nuevos y emocionantes enfoques de aprendizaje.