

ترجمه کامل مقاله:

An Overview of the Supervised Machine Learning Methods

چکیده (Abstract)

در دهه گذشته تعداد زیادی از روش‌های یادگیری نظارت‌شده در حوزه یادگیری ماشین معرفی شده‌اند. یادگیری نظارت‌شده به یکی از موضوعات اصلی تحقیقات در یادگیری ماشین تبدیل شده است و بسیاری از تکنیک‌های آن برای پردازش و تحلیل انواع داده‌ها به کار گرفته می‌شوند. یکی از ویژگی‌های مهم یادگیری نظارت‌شده، استفاده از داده‌های برچسب‌خورده است که شامل برچسب‌هایی برای فرایند طبقه‌بندی هستند. این مقاله جنبه‌های بنیادی چند روش نظارت‌شده را خلاصه می‌کند و هدف اصلی آن، ارائه یک مرور کلی از یادگیری ماشین و معرفی تکنیک‌های آن است.

مقدمه (Introduction)

یادگیری ماشین حوزه‌ای گسترده است که در فناوری اطلاعات، آمار، احتمال، هوش مصنوعی، روان‌شناسی، نوروبیولوژی و بسیاری از علوم دیگر استفاده می‌شود. در یادگیری ماشین، حل مسائل با ساختن مدلی انجام می‌شود که نماینده مناسبی از مجموعه داده انتخابی باشد. با پیشرفت فناوری‌های پردازشی و داده‌های عظیم (Big Data)، یادگیری ماشین امروزه بسیار متفاوت از گذشته است. امروزه الگوریتم‌های متعددی توسعه یافته، بهبود پیدا کرده و قادر به انجام محاسبات پیچیده روی داده‌های بزرگ هستند. برنامه‌نویسی تطبیقی نیز بسیار محبوب شده است؛ جایی که برنامه‌ها قادر به تشخیص الگوها، یادگیری از تجربه، استخراج اطلاعات جدید از داده‌ها و بهینه‌سازی دقت و کارایی خروجی خود هستند.

انواع یادگیری ماشین

بر اساس نوع خروجی مورد انتظار، الگوریتم‌های یادگیری ماشین به دسته‌های زیر تقسیم می‌شوند:

1. یادگیری نظارت‌شده: ایجاد تابعی که ورودی‌ها را به خروجی‌های مطلوب نگاشت کند. غالباً در مسائل طبقه‌بندی استفاده می‌شود.
2. یادگیری بدون نظارت: داده‌ها بدون برچسب هستند و هدف پیدا کردن ساختار پنهان میان داده‌هاست.

3. یادگیری نیمه نظارت شده:
- ترکیبی از داده‌های برچسب خورده و بدون برچسب.
4. یادگیری تقویتی:
- الگوریتم با تعامل با محیط و دریافت پاداش یاد می‌گیرد.
5. یادگیری انتقالی / استنتاجی (Transduction):
- بدون ساخت تابع صریح، خروجی‌های جدید را از ورودی‌های جدید و قدیمی حدس می‌زنند.
6. یادگیری برای یادگیری (Learning to Learn):
- الگوریتم از تجربیات قبلی، تمایلات استنتاجی خود را یاد می‌گیرد.

کارهای مرتبط (Related Work)

مقالات متعددی تکنیک‌های یادگیری ماشین و طبقه‌بندی را بررسی کرده‌اند.
از جمله:

- مقاله Caruana & Mizil که ۱۰ روش یادگیری نظارت شده (از جمله SVM، شبکه عصبی، لجستیک، بیز ساده و ...) را مقایسه می‌کند.
- پژوهش‌هایی درباره طبقه‌بندی اسناد متنی، تحلیل داده‌های پژوهشی، و مقایسه دقت، سرعت و پیچیدگی الگوریتم‌ها.
- کتاب‌ها و فصل‌هایی که درباره SVM، KNN، درخت تصمیم و سایر ابزارهای یادگیری نظارت شده توضیح می‌دهند.

پس زمینه: یادگیری نظارت شده (Background: Supervised Learning)

فرایند یادگیری نظارت شده شامل دو مرحله است:

1. آموزش (Training):
 - ویژگی‌ها توسط الگوریتم آموخته شده و مدل ساخته می‌شود.
 2. آزمون (Testing):
 - مدل روی داده‌های جدید اعمال شده و خروجی پیش‌بینی می‌شود.
- الگوریتم، داده‌های ورودی را همراه با برچسب‌های صحیح دریافت می‌کند و با مقایسه خروجی و مقدار واقعی، خطا را کاهش می‌دهد.

دو نوع اصلی وظایف نظارت شده

- طبقه‌بندی (Classification): خروجی گسسته است.
- رگرسیون (Regression): خروجی عددی و پیوسته است.

یادگیری نظارت شده بیشترین کاربرد را در شبکه‌های عصبی، درخت تصمیم و کاربردهای پیش‌بینی داده‌های تاریخی دارد (مثل پیش‌بینی نوع گل زنبق).

★ الگوریتم‌های یادگیری نظارت شده (Supervised Learning Algorithms)

(۱) درخت تصمیم (Decision Tree)

- درخت شامل گره ریشه، گره‌های داخلی و برگ‌هاست.
- هر گره یک آزمون روی ویژگی‌ها انجام می‌دهد.
- نمونه با عبور از گره‌ها به برگ مناسب می‌رسد.
- می‌تواند ویژگی‌های عددی و رده‌ای را پردازش کند.
- ساده، قابل فهم، اما مستعد بیش‌برازش.

(۲) رگرسیون خطی (Linear Regression)

- هدف: یافتن رابطه خطی بین متغیرهای ورودی و خروجی.
- خروجی پیوسته است.
- مدل با کمینه‌سازی تابع هزینه، خط مناسب را پیدا می‌کند.

(۳) بیز ساده (Naïve Bayes)

- مبتنی بر قضیه بیز.
- فرض استقلال ویژگی‌ها دارد.
- سریع، مقاوم در برابر نویز، و مناسب طبقه‌بندی متنی.

(۴) رگرسیون لجستیک (Logistic Regression)

- یک طبقه‌بند تمایزی (Discriminative) است.
- از تابع سیگموید استفاده می‌کند:

$$\frac{1}{z \cdot e + 1} = g(z)$$

- خروجی بین ۰ و ۱ است و احتمال رخداد یک کلاس را نشان می‌دهد.

۵) نزدیک‌ترین همسایه‌ها (KNN)

- نمونه جدید با توجه به K همسایه نزدیک خود طبقه‌بندی می‌شود.
- محاسبات زیاد دارد اما ساده و دقیق است.

۶) ماشین بردار پشتیبان (SVM)

- به دنبال یافتن بهترین ابرصفحه است که دو کلاس را از هم جدا کند.
- برای داده‌های با ابعاد بالا بسیار مناسب است.
- با کرنل‌ها قابلیت تفکیک داده‌های غیرخطی را دارد.

۷) جنگل تصادفی (Random Forest)

- مجموعه‌ای از درخت‌های تصمیم است که با داده‌ها و ویژگی‌های تصادفی ساخته می‌شوند.
- دقت بالا و مقاومت خوب در برابر بیش‌برازش دارد.

۸) AdaBoost

- چندین مدل ضعیف را با وزن‌دهی متفاوت ترکیب می‌کند.
- نمونه‌های اشتباه وزن بیشتری می‌گیرند تا مدل‌های بعدی روی آن‌ها تمرکز کنند.

۹) Gradient Boosting

- مدل‌ها به صورت مرحله‌ای ساخته می‌شوند و هر مدل خطای مدل قبلی را اصلاح می‌کند.
- دقت بسیار بالا، ولی حساس به بیش‌برازش.

۱۰) XGBoost

- نسخه سریع‌تر و بهینه‌تر Gradient Boosting.
- دارای Regularization (جلوگیری از بیش‌برازش).
- از بهترین الگوریتم‌ها در رقابت‌های Kaggle.

۱۱ (یادگیری عمیق (Deep Learning)

- شامل شبکه‌های عصبی چندلایه (CNN، RNN و ...).
- قدرت بسیار بالا در تحلیل تصویر، صوت و متن.
- نیازمند داده و توان پردازشی زیاد.

نتیجه‌گیری (Conclusion)

می‌توان نتیجه گرفت که یادگیری نظارت‌شده یکی از مهم‌ترین روش‌ها در یادگیری ماشین است. روش‌های نظارت‌شده معمولاً موفق‌تر از روش‌های بدون‌ناظر هستند زیرا داده‌های برچسب‌خورده معیارهای واضحی برای آموزش مدل فراهم می‌کنند. این مقاله مرور کوتاهی از فرایند یادگیری، ساختار الگوریتم‌ها و برخی از روش‌های رایج یادگیری نظارت‌شده ارائه داد. با پیشرفت سریع این حوزه، فضای زیادی برای توسعه و بهبود الگوریتم‌ها وجود دارد و انتظار می‌رود عملکرد آن‌ها در آینده حتی بهتر شود.