ارائه روشی خبره برای تولید درختهای تصمیم

فرید سیفی[1]، محمد رضا کنگاوری[2]

چکیده

تا کنون روشهای متعددی برای تولید درختهای تصمیمِ دسته بند ابداع شده اند که معمولا برای تصمیم گیری و پیش بینی بکار می روند. این روشها سعی در بهینه سازی پارامترهایی چون دقت، سرعت دسته بندی، اندازه درختهای ساخته شده، سرعت یادگیری و میزان حافظه بکار رفته دارند. بین پارامترهای ذکر شده تناقض وجود داردبدین معنی که بهینه سازی یک پارامتر ممکن است موجب تغییرات نامناسب سایر پارامترها شود، به همین دلیل است که تمام روشهای موجود سعی در ایجاد توازن بین این پارامتر ها دارند.

در این تحقیق- با در نظر گرفتن تاثیر تمام مجموعه داده های یادگیری بروی تخصیص کلاس به هر نمونه داده- روشی جدید برای ایجاد درختهای تصمیم ارائه کرده ایم که درختهایی با دقت نسبتا مناسب و با پیچیدگی بسیار کم را در زمانی بسیار کوتاه و با بکار گیری حافظه ای اندک تولید می کند. به منظور رسیدن به این هدف یک فرایند چند مرحله ای بکار برده ایم. در هر مرحله این فرایند مجموعه داده های یادگیری یکبار از ابتدا به انتها و بار دیگر در جهت عکس مورد بررسی قرار می گیرد تا الگوی کلاسها برای انتخاب متغیر استخراج شود. سپس با استفاده از متغیر منتخب- در هر مرحله- شاخه های جدید در درخت ایجاد می شوند. در پایان هر مرحله و پس از ایجاد شاخه های جدید در درخت، متغیر منتخب و تعدادی از نمونه داده ها از مجموعه داده های یادگیری حذف می شوند. این عملیات در مراحل مختلف و بصورت متناوب بروی داده ها و متغیر های باقی مانده ادامه می یابد تا زمانیکه درخت بطور کامل ساخته شود.

در این تحقیق مجموعه داده های شناخته شده ای که قبلا در تحقیقات مختلف بکار گرفته شده اند را بکار برده ایم و دقت و اندازه درخت ایجاد شده توسط این روش را با سایر روشهای مطرح مقایسه کرده ایم.

کلمات کلیدی

دسته بندی، درخت تصمیم، داده کاوی، تصمیم گیری، اکتشاف دانش، پیش بینی

Presentation of a Sophisticated Approach to Decision Tree construction

Farid Seifi, Mohammad Reza Kangavari

ABSTRACT

Decision tree classification is one of the most practical and effective methods which is used in inductive learning. Many different approaches, which are usually used for decision making and prediction, have been invented to construct decision tree classifiers. These approaches try to optimize parameters such as accuracy, speed of classification, size of constructed trees, learning speed and the amount of used memory. There is trade off between these parameters. That's to say that optimization of one may cause obstruction in the other, hence all existing approaches try to establish equilibrium.

In this study, considering the effect of the whole data set on class assigning of any data, we propose a new approach to construct not perfectly accurate, but less complex trees in a short time, using small amount of memory. For achieving this purpose, a multi-step process has been used. We trace the training data set twice in any step, from the beginning to the end and vice versa, to extract the class pattern for attribute selection. Using the selected attribute, we make new branches in the tree. After making branches, the selected attribute and some records of training data set are deleted at the end of any step. This process continues alternatively in several steps for remaining data and attributes until the tree is completely constructed.

In order to compare this new approach with previous ones we used some known data sets which have been used in different researches. Experimental results showed that it is efficient to use this approach particularly in cases of massive data sets, memory restrictions or short learning time.

KEYWORDS

Classification, Decision Tree, Data Mining, Decision Making, Knowledge Discovery, Prediction.

---

[1] دانشجوی کارشناسی ارشد، دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران. farid@comp.iust.ac.ir

[2] استادیار، دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران. kangavari@iust.ac.ir