ارایه و بررسی روشی موثرجهت انتخاب صفات خاصه مناسب در ساخت درخت تصمیم

مهدی اسماعیلی [۱] ؛ منصور طرفدار [۲]

چکیده

حذف و نادیده گرفتن صفات خاصه در پایگاه داده های حجیم جهت تصمیم گیری و تحلیل سریع می‌تواند موثر واقع شود. یکی از تکنیک های مهم و پر کاربرد در مراحل پیش پردازش داده‌ها ، کاهش ابعاد داده ها و یا به عبارت دیگر ، انتخاب صفات خاصه مناسب برای داده کاوی است.

در این مقاله تاثیر کاهش ابعاد در دسته بندی داده ها بررسی و پس از آن الگوریتمی موثر پیشنهاد می‌شود. الگوریتم از دو مرحله تشکیل شده است که در مرحله اول با کمک دو معیار، ترتیبی بین صفات خاصه به دست می‌آوریم. صفات خاصه مرتب شده در مرحله اول به عنوان ورودی برای مرحله بعدی که ساخت یک درخت تصمیم است استفاده می‌شود. نتایج نشان می‌دهند که درخت تصمیم حاصل از این الگوریتم پیشنهادی همراه با یک دقت قابل قبول کوچکتر خواهند بود.

کلمات کلیدی

درخت تصمیم، قوانین دسته بندی، کاهش ویژگی ها

Feature Selection as an Improving Step for Decision Tree Construction

Mehdi Esmaeili; Mansour Tarafdar

Abstract

The removal of irrelevant or redundant attributes could benefit us in making decisions and analyzing data efficiently. Feature Selection is one of the most important and frequently used techniques in data preprocessing for data mining.

In this work, special attention is made on feature selection for classification with labeled data. Here an algorithm is used that arranges attributes based on their importance using two independent criteria. Then, the arranged attributes can be used as input one simple and powerful algorithm for construction decision tree. Results indicate that this decision tree using featured selected by proposed algorithm outperformed decision tree without feature selection. From the experimental results, it is observed that, this method generates smaller tree having an acceptable accuracy.

Key Words

Decision Tree, Classification Rules, Features Reduction

۱. مقدمه

انتخاب صفات خاصه مناسب نقش مهمی در تکنیک های داده کاوی بازی می‌کند. بدون شک متدها معمولاً با داده های کاهش یافته سریع تر عمل می‌کنند. وجود داده های تکراری و نامربوط به طور معمول فقط به افزایش حجم داده ها و یا گمراهی متدها منجر می‌شود.

۱. عضو هیئت علمی دانشگاه آزاد اسلامی کاشان m.esmaeili@iaukashan.ac.ir

۲. دانشجوی کارشناسی ارشد نرم افزار دانشگاه آزاد اسلامی قزوین ، دانشکده برق رایانه و فناوری اطلاعات tarafdar.mansour@gmail.com

از سویی دیگر می‌دانیم که الگوریتم های دسته بندی یکی از متدهای رایج در داده کاوی به حساب می‌آیند. در این بین درخت های تصمیم به دلیل سادگی فهم از محبوبیت خوبی برخوردار هستند. هدف اصلی انتخاب صفات خاصه در الگوریتم های دسته بندی، کاهش صفات خاصه در ساخت درخت تصمیم است به صورتی که یک دقت قابل قبول برای مدل ساخته شده حفظ شود. برای یافتن حالت بهینه از ترکیب صفات خاصه، کلیه حالات آن باید بررسی و تحلیل شوند. روشن است که در بسیاری از مواقع این بررسی بسیار پیچیده بوده و یا حتی شدنی نیست.

تحت شرایط خاصی می‌توان از الگوریتم هایی بهره برد که نیمه بهینه[۱] هستند. هرچند که از نام این الگوریتم ها مشخص است این روش ها بهترین گزینه را انتخاب نمی‌کنند اما با انتخاب زیر مجموعه ای مناسب از صفات خاصه و پیچیدگی کمتر، کارایی بهتری از خود نشان می‌دهند. جستجو برای انتخاب صفات خاصه می‌تواند به صورت اتفاقی شروع شود (استفاده از الگوریتم هایی نظیر الگوریتم ژنتیک)[۱].

شاخص های ارزیابی از نقطه نظر الگوریتم هایی که پس از انتخاب صفات خاصه بر روی داده ها اجرا می‌شوند به دو دسته مستقل[۲] و غیرمستقل[۳] دسته بندی می‌شوند[۲]. یک معیار مستقل سعی می‌کند تا زیر مجموعه انتخابی از صفات خاصه[۴] را با کمک خصوصیات و خواص خود داده ها بدست آورد بدون اینکه به الگوریتم اجرایی در مراحل بعد توجهی داشته باشد. برخی از پر کاربرد ترین آنها عبارتند از: معیارهای مبتنی بر فاصله[۵]، معیارهای اطلاعاتی[۶] و معیارهای سازگاری[۷][۳][۴][۵][۶]. این درحالی است که معیارهای غیر مستقل کاملاً وابسته به الگوریتم اجرایی بعدی هستند. بدین صورت که به منظور ارزیابی روش پس از انتخاب زیر مجموعه ای از صفات خاصه، کارایی الگوریتم با کمک این زیرمجموعه انتخابی ارزیابی می شود و مناسب ترین زیر مجموعه، وابسته به اجرای بهتر الگوریتم با این صفات خاصه خواهد بود.

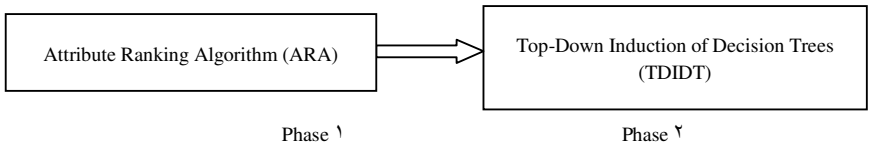
اخیراً تحقیقات زیادی بر روی انتخاب صفات خاصه مناسب جهت دسته بندی داده ها انجام شده است. لیستی از آنها در [۱][۲][۷][۸] آمده اند.

مقاله حاضر از الگوریتمی جهت رتبه بندی[۸] صفات خاصه براساس اهمیت شان استفاده می‌کند. این الگوریتم از دو معیار مستقل بهره می‌برد. پس از آن صفات خاصه به ترتیب اهمیت شان وارد مرحله ساخت درخت تصمیم خواهند شد.

این مقاله به صورت زیر سازماندهی شده است: بخش دوم جزئیات الگوریتم پیشنهادی را توضیح می‌دهد. در بخش سوم خصوصیات مجموعه داده هایی که برای ارزیابی الگوریتم پیشنهادی انتخاب شده‌اند را به صورت خلاصه آورده‌ایم. بخش چهارم شامل نتایج حاصل از الگوریتم پیشنهادی و مقایسه آن با چند الگوریتم دیگر است و در بخش پایانی نتیجه گیری و پیشنهاداتی را برای کارهای آینده بیان کرده ایم.

۲. روش پیشنهادی

روش پیشنهادی دارای دو مرحله است که در شکل ۱ نمودار اجمالی آن نشان داده شده است.



شکل ۱ : دیاگرام روش پیشنهادی

۲,۱. مرحله اول : رتبه بندی صفات خاصه

در این مرحله همانطور که از نام آن مشخص است از الگوریتمی (ARA) جهت بررسی اهمیت صفات خاصه استفاده می‌شود. بدین منظور از معیارهای سنجشی که در [۹] به آن اشاره شده است استفاده کرده ایم.

در الگوریتم ARA ما به دنبال صفات خاصه ای می‌گردیم که با کمک آن می توان تا حد امکان دو کلاس را از یکدیگر تمیز داد. همراه با این موضوع از معیار دیگری جهت ارزیابی همبستگی[۹] صفت خاصه به کلاس مربوطه نیز استفاده می شود. افزایش و بزرگی در این معیارها اهمیت بیشتر صفت خاصه را نشان می دهد به صورتی که می‌توان اهمیت یک صفت خاصه را از برآورد مجموع این دو معیار ارزیابی نمود.

ورودی الگوریتم مجموعه داده هایی است با تعداد n صفت خاصه در C کلاس قرار گرفته اند. با حذف صفت خاصه اول اهمیت این صفت خاصه محاسبه می شود. این کار برای n صفت خاصه تکرار می شود. عدد بزرگتر نشان دهنده اهمیت بالاتر صفت خاصه است. بدین ترتیب توانسته ایم ترتیبی برای صفات خاصه در نظر بگیریم. فرمول های ۱ و ۲ نشان می دهند که این کار چگونه انجام می‌شود.

$$\text{Distance1} = \sum_{i=1}^{c} P_i \sum_{k=1}^{n_i} \left[(X_{ik} - m_i)(X_{ik} - m_i)^T \right]^{1/2} \qquad (۱)$$

$$\text{Distance2} = \sum_{i=1}^{c} P_i \left[(m_i - m)(m_i - m)^T \right]^{1/2} \qquad (۲)$$

در این فرمول ها P_i احتمال کلاس i ام را نشان می‌دهد. X ها مقادیر داده ها را نشان می دهند که قبل از هر گونه پردازشی نرمال سازی می‌شوند. X_{k_i} مقدار نرمال سازی شده k امین صفت خاصه از i امین نمونه (رکورد) را نشان می دهد. m_i و m به ترتیب بردارهایی هستند که میانگین را برای نمونه های کلاس i ام و میانگین کل مجموعه داده ها را مشخص می کنند. n_i تعداد نمونه ها در کلاس i است. به صورتی که می توان نوشت: $n = n_1 + n_2 + \ldots + n_c$

به منظور محاسبه همبستگی صفت خاصه k ام و کلاس ها نیز از فرمول ۳ استفاده می کنیم:

$$\text{Attribute class correlation} = \sum_{i \# j} \left| X_{ik} - X_{jk} \right| \qquad (۳)$$

این فرمول برای صفات خاصه ای که متعلق به یک کلاس نیستند محاسبه می‌شود.

۲,۲. مرحله دوم: ساخت درخت تصمیم

در این مرحله از یک الگوریتم ساده در عین حال قدرتمند جهت تولید شروط[۱۰] درخت تصمیم استفاده می شود. شکل ۲ این الگوریتم را نشان می دهد.

```
IF all the instances in the training set belong to the same class THEN
    Return the value of class
ELSE (a) Select an attribute A from ranked list
     (b) Sort the instances in the training set into subsets,
         one for each value of attribute A
     (c) Return a tree with one branch for each non-empty
         subset,Each branch having a descendant subtree or
         a class value Produced by applying the algorithm
         recursively
```

شکل ۲ : الگوریتم ساخت درخت تصمیم

۳. توصیف داده های تحت آزمایش

از آنجا که معمولا روش‌های آماده‌سازی داده‌ها وابسته به نوع کاربرد هستند در این قسمت سعی بر آن شده تا داده‌هایی با خصوصیات متفاوت انتخاب شوند. این مجموعه داده‌ها از پایگاه داده UCI انتخاب شده‌اند [۱۰]. در جدول ۱ این چهار مجموعه داده ها به صورت خلاصه توصیف شده اند.

جدول ۱ : مشخصات کلی از دادههای تحت آزمایش

	of Number Attributes	Number of Instances	Number of Classes
Iris	٤	١٥٠	٣
Monk's Problems	٧	٤٣٢	٢
Glass Identification	١٠	٢١٤	٦
Ionosphere	٣٤	٣٥١	٢

۴. نتایج تجربی و بحث

در این بخش خروجی الگوریتم پیشنهادی را بر روی چهار مجموعه داده های معرفی شده در بخش قبل بررسی می کنیم. خروجی فاز اول الگوریتم برای مجموعه داده ها در جدول ۲ نشان داده شده است. در این جدول رتبه بندی میان صفات خاصه هر مجموعه داده را نشان می‌دهد.

جدول ۲ : رتبه بندی صفات خاصه (خروجی فاز اول)

Data Set	Attributes Ordering
Iris	۳,۲,۱,٤
Monk's Problems	۲,٥,٤,۱,۳,٦
Glass Identification	۲,۳,۸,٤,۱,٦,۷,٥,۹
Ionosphere	١٤,٢٠,٢٢,١٢,٣٠,٤,٢٧,٢٨,١٨,١٦,٢٤,٢,٦,١٠,٢٣,٣٢,٧, ١٩,٨,٣١,٢١,٢٥,١٧,١٣,٢٩,١١,٢٦,٩,٣,٥,١٥,٣٣,٣٤,١

همانطور که مشاهده می کنید در مجموعه داده Iris به ترتیب صفت خاصه سوم و دوم و اول و چهارم اهمیت صفات خاصه را در این مجموعه داده تشکیل می دهند. شکل ۳ خروجی فاز دوم الگوریتم پیشنهادی را بر روی نمونه داده Iris نشان می دهد.

```
Field۳<=۱.۷ : Iris-setosa( ٤٨)
Field۳>۱.۷ & Field۲<=۲.۲ : Iris-versicolor(٤/۱)
Field۳>۱.۷ & Field۲>۲.۲ & Field۱<=٤.۹ : Iris-versicolor(۳/۲)
Field۳>۱.۷ & Field۲>۲.۲ & Field۱>٤.۹ & Field٤<=۱.٤ : Iris-versicolor(۳٤/۲)
Field۳>۱.۷ & Field۲>۲.۲ & Field۱>٤.۹ & Field٤>۱.٤ : Iris-virginica(٦۱/۱٤)
```

شکل ۳ : خروجی فاز دوم الگوریتم برروی داده Iris

چنانچه به پنج شرط تولید شده توسط برنامه توجه کنید خواهید یافت که ترتیب آنها همان ترتیب اهمیت صفات خاصه است که در فاز اول الگوریتم به دست آمده است. اعداد داخل پرانتز نشان دهنده تعداد نمونه هایی است که شامل شروط مربوط می شوند. جدول ۳ صحت و دقت روش پیشنهادی بر روی چهار مجموعه داده انتخابی نشان می دهد.

جدول ۳ : محاسبه پارامترها به تفکیک کلاس

	TP Rate	FP Rate	Recall	Precision	F-measure	Class
Iris	۰.۹٦	۰	۰.۷۲	۱.۰۰	۰.۸٤	Iris Setosa
	۰.۷۲	۰.۰٥	۰.۷۲	۰.۸۸	۰.۸۰	Iris Versicolour
	۰.۹٤	۰.۱٤	۰.۹۰	۰.۷۷	۰.۸۳	Iris Virginica
Monk's Problems	۰.٦۷	۰.۳۳	۰.٦۷	۰.٦۷	۰.٦۷	Class ۰
	۰.٦۷	۰.۳۳	۰.٦۷	۰.٦۷	۰.٦۷	Class ۱
Glass Identification	۰.۱۰	۰.۰۳	۰.۰۷	۰.٦٤	۰.۱۳	Building_w_f_p
	۰.۹٦	۰.٥۸	۰.۸٤	۰.٤۸	۰.٦۱	Building_w_nf_p
	۰.۰٦	۰.۰۱	۰.۰۱	۰.۳٤	۰.۰۲	Vehicle_w_f_p
	۰.٥٤	۰.۰۱	۰.۰۷	۰.۷۸	۰.۱۳	Containers
	۰.۷۸	۰.۰۲	۰.۰۸	۰.٦٤	۰.۱٤	Tableware
	۰.۸٦	۰.۰۱	۰.۲۲	۰.۹۳	۰.۳٦	Headlamps
Ionosphere	۰.۷۰	۰.۱۰	۰.۷۰	۰.۸۷	۰.۷۸	Bad
	۰.۹٤	۰.۱۷	۰.۹٤	۰.۸٥	۰.۸۹	Good

از ستون آخر جدول ۳ روشن است این مقادیر به صورت مجزا برای هر کلاس محاسبه شده است.

بدون شک مدلهای تولید شده توسط الگوریتم های داده کاوی هرچه کوچکتر باشند قابل فهم تر خواهند بود. در دسته بندی داده ها و به خصوص درخت های تصمیم تعداد شروط استخراج شده توسط الگوریتم یکی از مهمترین پارامترهایی به شمار می رود که در مقایسه کارایی الگوریتم ها بر روی آن تاکید می شود. در جدول ۴ مقایسه ای میان چهار الگوریتم و متد پیشنهادی براساس تعداد شروط تولید شده و اندازه درخت انجام شده است.

جدول ۴ : تعداد شروط تولیدشده و اندازه درخت توسط الگوریتمها

	J٤٨	BFTree	REPTree	NBTree	Proposed Method
Iris	٥/٩	٦/١١	٣/٥	٤/٧	٥/٨
Monk's Problems	٢/٣	٢/٣	٨/١٥	١/١	٤/٦
Glass Identification	٣٠/٥٩	١٦/٣١	١٢/٢٣	٩/١٧	١٤/٣٠
Ionosphere	١٨/٣٥	١١/٢١	٥/٩	٨/١٥	١٣/٢٤

همانطور که از جدول مشخص است روش پیشنهادی در بین این روشها بهترین نیست اما عکس العمل آن نسبت به نوع داده ها خوب است. توجه کنید که ما در ساخت درخت از هیچ راهبرد اکتشافی[11] استفاده نمی‌کنیم و این درحالی است که دیگر الگوریتم ها پیچیده تر از فاز دوم الگوریتم ما عمل می کنند. در جدول ۵ نرخ خطای روش پیشنهادی نیز بررسی شده است.

جدول ۵ : مقایسه نرخ خطای الگوریتم ها

	Error Rate				
	J٤٨	BFTree	REPTree	NBTree	Proposed Method
Iris	٠.٠٤	٠.٠٦	٠.٠٦	٠.٠٦	٠.١٢
Monk's Problems	٠.٢٥	٠.٢٥	٠.١٥	٠.٢٥	٠.٣٣
Glass Identification	٠.٣٤	٠.٣٣	٠.٣٨	٠.٣٠	٠.٤٥
Ionosphere	٠.٠٩	٠.١٠	٠.١١	٠.١٠	٠.١٨

۵. نتیجه گیری

در این مقاله تاثیر مرحله پیش پردازش و آماده سازی داده ها در ساخت یک درخت تصمیم بررسی شد. انتخاب صفات خاصه مناسب برای ساخت یک درخت تصمیم نقش مهمی بازی می‌کند. در الگوریتم پیشنهادی ابتدا ترتیبی میان صفات خاصه در داده ها به دست می‌آید. همبستگی داده‌ها به برچسب کلاس ها و تغییر مقادیر دادهها با تغییر کلاس از جمله پارامترهایی است که در اهمیت رتبه بندی یک صفت خاصه محاسبه می‌شود. در مرحله بعدی با توجه به این رتبه بندی درخت تصمیم و یا در واقع شروط ایجاد می‌شوند. نتایج به دست آمده نشان می دهد درختی که توسط روش پیشنهادی تولید می شود به طور متوسط کوچکتر از درخت هایی است که توسط چهار متد انتخابی ایجاد شده است.

چنانچه مایل باشیم تا درخت حاصل از الگوریتم پیشنهادی دارای دقت بالاتر و بهتری باشد می‌توانیم مرحله اول الگوریتم را پس از حذف صفت خاصه به صورت تکراری انجام دهیم. در این صورت الگوریتم از پیچیدگی بیشتری برخوردار و در نتیجه سرعت اجرای آن کاهش می‌یابد.

بهرحال به عنوان کار آینده می‌توان افزودن یک روش هرس کردن مناسب برای درخت را آزمایش نمود. بدون شک چون درخت تولید شده توسط روش پیشنهادی ترتیبی در تست کردن شروط دارد (Oblivious Tree) شاید بتوان قوانین مناسب تر و بهتری برای هرس کردن آن با حفظ دقت مناسب جستجو نمود.

۶. مراجع

[۱] J.Doak , *An Evaluation of Feature Selection Methods and Their Application to Computer Security* , technical report, University of California at Davis , Department of Computer Science, ۱۹۹۲

[۲] H.Liu, L.Yu , *Toward Integrating Feature Selection Algorithms for Classification and Clustering* , IEEE Transactions on Knowledge and Data Engineering , Vol. ۱۷, No. ۴, pp. ۴۹۱-۵۰۲, April-۲۰۰۵

[۳] H.Almuallim , T.G. Dietterich , *Learning Boolean Concepts in the Presence of Many Irrelevant Features* , Artificial Intelligence, Vol. ۶۹, pp.۲۷۹-۳۰۵, ۱۹۹۴

[۴] M.Ben-Bassat, *Pattern Recognition and Reduction of Dimensionality* , Handbook of Statistics-II, P.R. Krishnaiah and L.N. Kanal, pp. ۷۷۳-۷۹۱, North Holland, ۱۹۸۲

[۵] M.A.Hall , *Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning* , Proc. ۱۷th Int'l conf. Machine Learning, pp. ۳۵۹-۳۶۶, ۲۰۰۰

[۶] H.Liu , H.Motoda , Feature *Selection for Knowledge Discovery and Data Mining. Boston* , Kluwer Academic, ۱۹۹۸.

[۷] M.Dash , H.Liu , *Feature Selection for classification* , Intelligent Data Analysis: An Int'l J., Vol. ۱, No. ۳, pp. ۱۳۱-۱۵۶, ۱۹۹۷

[۸] W.Siedlecki , J.Sklansky , *On Automatic Feature Selection* , Int'l J. Pattern Recognition and Artificial Intelligence, Vol. ۲, pp. ۱۹۷-۲۲۰, ۱۹۸۸.

[۹] L.Wang, X.Fu, *Data Mining with Computational Intelligence* , Springer , pp. ۱۱۷-۱۲۳,۲۰۰۵ .

[۱۰] C.L. Blake, C.J.Merz, *UCI Repository of Machine Learning Databases*. Irvine, CA: University of California, Department of Information and Computer Science , (۱۹۹۸). [http://www.ics.uci.edu/~mlearn/MLRepository.html]

زیرنویس ها

[۱] Suboptimal

[۲] Independent

[۳] Dependent

[۴] Feature Subset

[۵] Distance Measures

[۶] Information Measures

[۷] Consistency Measures

[۸] Ranking

[۹] Correlation

[۱۰] Rules Generation

[۱۱] Heuristic approach