dataacademy.ir

روش های ایجادپراکندگی و استفاده ازخصیصه های متعدد درخوشه بندی ترکیبی

سوسن حسینی

دانشجوی کارشناسی ارشد،نرم افزارکامپیوتر،دانشگاه آزاداسلامی مشهد

Sosan.hosseini@gmail.com

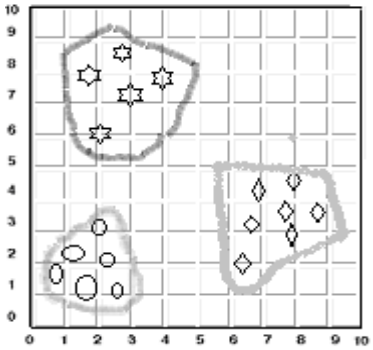
چکیده

خوشه بندی یکی از تکنیک های داده کاوی می باشدوبه عنوان مهم ترین مسأله دریادگیری بدون نظارت درنظرگرفته می شود. با ترکیب خوشه هامی توان به خوشه بندی ترکیبی برای دست یافتن به نتیجه ای، پایدار وقدرتمنددست یافت .و با انتخاب یک روش ایجادپراکندگی مناسب به خوشه بندی اصولی دست یافت.

کلید واژه- خوشه بندی،داده کاوی، ایجادپراکندگی،خصیصه های متعدد،خوشه بندی ترکیبی.

۱- مقدمه

یکی ازتکنیک های مهم داده کاوی،خوشه بندی است که داده های مشابه ومتفاوت راازهم جدامی کند.خوشه بندی به معنای تقسیم بندی بدون نظارت است که درآن داده ها به دسته هایی که از نظر پارامترهای مورد علاقه، شباهت بیشتری به یکدیگر دارند،تقسیم میگردند.ایده اصلی[1]درخوشه بندی اطلاعات جداکردن نمونه هاازیکدیگرو قراردادن آنهادرگروه های شبیه به هم می باشد.درشکل ۱ نمونه ای ازخوشه بندی دیده می شود.



شکل ۱:نمونه خوشه بندی[2]

هریک ازالگوریتم های خوشه بندی با توجه به اینکه بر روی جنبه های متفاوتی ازداده هاتأکید می کنند،داده ها را به صورت های مختلفی طبقه بندی می کند.به همین دلیل نیازمندشناخت وبررسی روش های ایجادپراکندگی درخوشه بندی ترکیبی هستیم.ادامه این مقاله اینگونه سازماندهی شده است:دربخش دوم مروری برخوشه بندی ترکیبی ومراحل آن ودربخش سوم به بررسی روش های ایجادپراکندگی ودربخش چهارم استفاده ازخصیصه های متعددرخوشه بندی ودربخش آخربه نتیجه گیری پرداخته خواهدشد.

۲- خوشه بندی ترکیبی

هریک از الگوریتم های خوشه بندی باتوجه به اینکه برروی جنبه های متفاوتی ازداده هاتأکیدمی کند,داده ها رابه صورت های متفاوتی طبقه بندی می نماید.درواقع هدف اصلی[3]خوشه بندی ترکیبی جستجوی بهترین خوشه ها باستفاده ازترکیب نتایج الگوریتم های دیگراست.خوشه بندی داده به صورت ترکیبی میتواندجواب های ایمن ترومؤثرتری راازنظرسادگی،پایداری وانعطاف پذیری ارایه دهد.مطالعات اخیر[4]نشان می دهدکه اجماع خوشه بندی می تواندخارج ازوضعیت های نوع رأی گیری بااستفاده ازروش های

آکادمی داده
dataacademy.ir

dataacademy.ir

مبتنی برگراف وتئوری اطلاعات بدون حل دقیق مشکل تناظربرچسب ها انجام شود،روش خوشه بندی ترکیبی،باترکیب افزارهای مختلف تولیدشده خوشه بندی پایه یک افزارمستحکم ازداده راتولید می کند.دراکثرالگوریتم های پایه برای خوشه بندی ترکیبی ازنمونه برداری داده هااستفاده می شود.مسأله اصلی دراین روش هاچگونگی ارزیابی خوشه وافراز است.درچندسال اخیر[5] پایداری خوشه به عنوان یک معیارارزیابی خوشه موردتوجه قرارگرفته است.

۱- ۲ مراحل خوشه بندی ترکیبی

دومرحله مهم درخوشه بندی ترکیبی وجوددارد، که عبارتنداز[6]:

۱- ترکیب نتایج بدست آمده ازخوشه بندی های متفاوت اولیه برای تولیدخوشه بندی نهایی که این کارتوسط تابع توافقی (الگوریتم های ترکیب کننده) انجام می گردد.

۲- تولیدنتایج مختلف ازخوشه بندی هابه عنوان نتایج خوشه بندی اولیه براساس اعمال روش های مختلف که این مرحله ایجادتنوع یاپراکندگی تعریف می شوند،که درادامه بررسی می گردد.

۲- ۲ تابع توافقی

ترکیب نتایج بدست آمده ازخوشه بندی های متفاوت اولیه برای تولیدخوشه نهایی برای پیداکردن افزارهای ترکیب شده تعریف می شوند.به طورکلی این روش ها،روش مبتنی برماتریس-رأی گیری-تئوری اطلاعات وروش مبتنی بر،ابرگراف ومدل آمیزشی نامیده می شوند.اکثرمطالعات اخیر[7]درحوزه خوشه بندی ترکیبی سعی دارند،ابتداخوشه بندی های اولیه تاحدممکن پراکنده تولید شوند وسپس بااعمال یک تابع توافقی همه این نتایج باهم ترکیب شوند.

۲-۳ ایجادتنوع

تنوع وپراکندگی باعث داشتن نتیجه مطلوب تردرخوشه بندی می گردد.که برای،بدست آوردن نتایج متنوع وویژگی های خاص ازداده ها،می توان ازدوروش کلی [8] استفاده نمود:۱- استفاده ازالگوریتم های خوشه بندی.۲- استفاده ازروش های مختلف ایجادپراکندگی برروی تنهایک الگوریتم خوشه بندی پایه وتقسیم بندی داده به قسمت های متفاوت. شامل زیرمجموعه های مختلف ازداده وویژگی که درادامه به آن خواهیم پرداخت.

۳- بررسی روش های ایجادتنوع وپراکندگی

اکثرمطالعاتی[9] که تاکنون درزمینه خوشه بندی ترکیبی انجام گرفته است،جستجوی روش هایی است که بتواندپراکندگی نتایج مجمع رده بندی هارابهترکند.درادامه به بررسی سه روش ایجادتنوع وپراکندگی پرداخته خواهدشد.

۳-۱تغییرمقادیراولیه خوشه بندی:

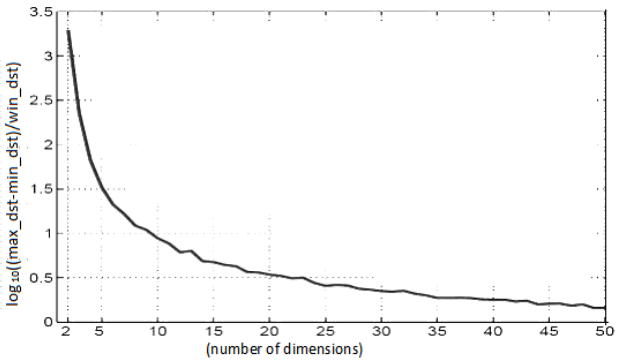
یکی ازراههای الگوریتم افزایش پراکندگی تغییر پارامتر های اولیه الگوریتم های خوشه بندی می باشد.دراین روش میتوان تنهاازیک الگوریتم خوشه بندی استفاده نمود.انتخاب اولیه مراکزخوشه ها می تواندبه دوصورت تصادفی وهوشمندانه انجام گردد.

۳-۲ انتخاب زیرمجموعه مختلف ازویژگی ها:

این روش برروی زیرمجموعه های مختلف ازویژگی هااجرامی شود. ونشان داده شده است،که زیرمجموعه بهترمنجربه خوشه پذیری،بهتری میگردد.درشکل2 می بینید،وقتی ابعادزیادمی گردد واسط بین نقاط می گرددوهمچنین کاهش ابعادمنجربه تجسم بهترداده وکاهش زمان ومقدار حافظه می گردد.

آکادمی داده
dataacademy.ir

dataacademy.ir



شکل۲:مقایسه رابطه واسط نقاط وابعاد[10] .

۳-۳ انتخاب زیرمجموعه های مختلف ازداده:

یکی،ازمهمترین عواملی که در بالابردن کیفیت نهایی خوشه بندی تاثیردارد، ایجادپراکندگی بیشترمی باشد،که با انتخاب تعدادمحدودی ازنمونه ها[11] به جای کل نمونه هاباعث کاهش میزان محاسبات وافزایش پراکندگی می گردد.

٤- استفاده ازخصیصه های متعدد

درخوشه بندی درجات متفاوتی از شباهت میان داده ها ومراکز ثقل انتخاب شده وجوددارد.اگرهرالگودقیقاً بایک مرکز منطبق شود، تفکیک قطعی یاسخت امکان پذیر است وداده ها رامی توان درخوشه های مجزا دسته بندی کرد.اما دراغلب موارد مرزهای بین خوشه هاخوش تعریف نیستند؛لذاالگوهارادرجات تعلق مختلف به خوشه ها نسبت می دهیم.یکی از انواع خوشه بندی فازی که کاربردزیادی داردخوشه بندی عددی fcm (fuzzy c-mean)که دراصل برای خوشه بندی داده باخصیصه های عددي ارایه شده است.درfcmمعیارعدم تشابه بین داده ها یک فاصله اقلیدسی یافاصله اي که روی داده ها ی عددی دارای خصیصه های گروهی هستندتعریف مي شود.خوشه بندي داده هاي گروهي موضوع مهمي درداده کاوي وشناسايي الگو است. الگوريتم k-mode توسعه الگوي k-means براي خوشه بندي داده هاي گروهي است که در آن از يك معيار تطابق عدم تشابه ساده براي اشياء گروهي استفاده مي شود.دراين روش ازmode بجايmean استفاده مي شود.خصيصه هاي ترتيبي رامي توان بااستفاده از نگاشت هاي ترتيبي- عددي بصورت الگوريتم هاي خوشه بندي عددي در نظر گرفت. يادگيري تطبيقي نگاشت ترتيبي-عددي،در[12]رامي توان بعنوان بهبود كيفيت خوشه بندي فازي براي اشيائي با خصيصه هاي فازي مورد استفاده قرار داد.بهبود ديگري در معيار عدم تشابه بين اشياء و الگو ها بطور مجزا در انواع مختلف الگوريتم KM در[13]براي اشياء گروهي پيشنهاد شده است.دراين الگوريتم، يك الگوي خوشه بعنوان ليستي از mode ها در نظر گرفته شده است.يك mode در يك خوشه براي يك خصيصه گروهي خاص، ليستي از همه برچسبها در خصيصه به همراه رخدادهايشان در خوشه مي باشد.معيار عدم تشابه بين يك شئ والگو در سطح خصيصه،يك منهاي تعداد رخدادهاي مقدار خصيصه (برچسب) شئ در نظر گرفته مي شودكه الگوريتم FCM با الگو مستقل در[14] ارائه شده است. اگرچه همگرايي اين دو الگوريتم در [15] به اثبات نرسيده است، با اين وجود در كاربردواقعي،مي توان از معيار عدم تشابه مختلفي استفاده كرد.درجدول (۱)مجموعه هایA,B,C به همراه مشخصات اشان درج شده است.

آکادمی داده
dataacademy.ir

dataacademy.ir

نام شیء	تعداد هر شیء	خصوصیت عددی یاغیرعددی	ویژگی گروهی	ماکزیمم برچسب خصوصیت گروهی	تعداد کلاس
A	47	0	35	7	4
B	101	1	15	2	7
C	653	6	9	14	2

جدول ۱ : مشخصات سه مجموعه از اشیاء واقعي.[16]

الگوريتم GFCM، ۲۰۰بارروي داده هابامقادير مختلف p وm اجراشده است ومفروضات زيردرنظرگرفته شده است:1-درهراجرا،الگو هاي مختلفي بكارگرفته شده است.(مقدار $\varepsilon = 0.0001$براي شرط خاتمه درنظر گرفته شده است).-2اگرتعدادبرچسب يك خصيصه گروهي خاص،كمترازpباشد،تعداددقيق برچسب به جايpاستفاده مي شود.باتوجه به تعدادخوشه بكاررفته،ميانگين مقادير انديس براي همه مقاديرpاندازه ودرجدول(1-۱) نشان‌داده،شده است.

C: (WITH M= 1,1)									
MPC	6	6	9 ,10 ,13	13	14	11	12	11 ,12	11 ,13
CPI	6	6 ,7	9 ,10 ,13	9 ,11	14	11	12	11	11
XB	11	13	12	11	11	13 ,14	10	11	11
TANC	9	13	10	11	11	13	10	11	11
COC	6	6	7 ,11 ,12	13	7	14	12 ,14	11	11 ,13

جدول۱-۱:بهترين مقادير p با توجه به انديس هاي صحت خوشه فازي با تعداد متفاوت خوشه.[16] .

مقادير كوچك m بويژه (۱,۱)=m براي اشيائي با خصيصه هاي گروهي ، به خوشه بندي دقيق تري منجر مى گرددو بيشينه مقادير p براي A,B,C به ترتيب ۷ ، ۲ و ۱٤مي باشد.ضرايب فازي سازي متفاوتي براي m در نظر گرفته شده اند ، كه اين منجر به حصول مقادير انديس متفاوتي شده است.

٥-نتيجه گيرى

درخوشه بندى تركيبى وايجادپراكندگى درخوشه هابراى داشتن يک خوشه بندى تركيبى مؤثرترمطالعات ونتايج بسيارى مى توانندانجام پذيرد.مثلا مطالعاتى درزمينه بهبودانتخاب هوشمندانه مقاديراوليه الگوريتم هاوهمچنين انتخاب زيرمجموعه اى بهينه وكامل ازداده هاوويژگى ها براى ايجادتنوع وپراكندگى مفيدمى باشد. وبكارگيري الگو هاي تك خصيصه اي و نيز متدهاي تطابق عدم تشابه ساده در الگوريتم هاي خوشه بندي سنتي براي خصيصه هاي گروهي ، مي تواند به نتايج خوشه بندي غير دقيق منجر شود. در [17] مفهوم جديد الگوهاي خوشه p-mode فازي تعريف شده است كه شامل چندين برچسب در سطح خصيصه گروهي مي باشد،كه همگرايي اين

آکادمی داده
dataacademy.ir

dataacademy.ir

الگوريتم جديد (GFCM) به اثبات رسيده است. بررسي ها روي مجموعه اشياء واقعي نشان داده است كه مقادير p در الگوهاي p-mode فازي و ضرايب فازي سازي در كارايي الگوريتم FCM موثرندودر صورت بكارگيري متد تطابق عدم تشابه ساده براي مقاديرگروهي ، ضرايب فازي سازي كوچك ،مثل(۱,۱) روي اشيائي با خصيصه هاي گروهي منجر به خوشه بندي بهتري مي شوند. همچنين اگر p=max اختيار شود ، نتايج بهتري نسبت به(p =1) حاصل خواهد شد،البته اين امر در همه موارد صادق نيست.

مراجع

[1] F.kovas,c.legany,A.Babos,"cluster validity measurement techniqes",department of automation and applied informatics,Budapest university of technology and economics,pp.2-3,2003.

[2] Q.He,"A Review of clustering Algorithms as Applied inIR",Gr-aduate school of library and information science university of ilinois at urbana-champaign, ,pp97.-98,1999.

[3] Dudoit and fridly "bagging to improve the accurancy of a clustering procedure ",bioinformatics,pp.1090-1091,2003.

[4] Topchy,a.k and punch w.f"a mixture model for clustering ensemble",in proc.siam int1.conf.on data minig,sdm ,pp.389-390,2009.

[5] Fischer,b and buhmann,j.m"path-based clustering for grouping of smoth curves and texture segmentation",ieee trans.on pami,vol.25,No.4,pp.513-514,2003.

[6] Feyrn,x and brodle,c"random projection for high dimensional data clustering:a cluster ensemble approach ",in proc.20th int.conf.on machine learnimg,icml,pp.14-16, 2005.

[7] M. Halkidi, Y. Batistakis and M. Vazirgiannis, "On Clustering Validation Techniques", Journal of Intelligent Systems, vol.23,No.7, pp 107-145, 2001.

[8] Minaei-bidgoli,B.,jain,A.K and punch,W.F"a comparision of resampling method for clustering ensemble",in proc.int1.conf.machine learning methods technology and application ,Mlmta 04,las vegas, vol.37,No.6,pp 150-154,2006.

[9] Topchy,A.Minaei-Bidgoli,B.Jain,A.K.and punch,W.F"adaptive clustering ensemble",in proc.int1.conf on pattern recogation,ICPR,Cambridge,UK,pp.147-149,2004.

[10] Levine.e and domany.E,"Resampling metgod for unsupervised estimation of cluster validaty".neural Computation, vol.17,No.13,pp.2583-2593,2001.

[11] Javad azimi,Reza Davoodi,Morteza analoui ,"fast Convergence Clustring ensemble,In Conference of Data mining and Data warehouse",pp.14-16,2006.

[12] W.BOCK.Day,"consensus methods as tools for data analysis.in,editor,classification and related method for data analysis",Elsiver scince publishers B.V ,pp.317-320,1988.

[13] Fayyad u,Piatetsky-shapiro ,"Advances in know ledge discovery and data mining",pp.456-459,1996.

[14] Brunk.c,Kelly.j,"an integrated system for data mining ",proc.3 ,int.1.conf.on knowledge discovery and data minig,pp.135-138,1997.

[15] Becker.ra,chambers.jm,"an in teractive environment for data analysis and graphics ",pp.143-145,1984.

آکادمی داده
dataacademy.ir

dataacademy.ir

[16] M.Lee,w.pedrycz,”the fuzzy C-means algorithm with fuzzy p-mode protypes for clustering object s having mixed features”,pp.15-17,2009.

[17] Stefano Benati,”categorical data fuzzy clustering:an analysis of local search heuristics,Computers &Operations research”,pp.766-775,2008.

آکادمی داده
dataacademy.ir