تحلیل فرآیند داده کاوی مبتنی برخوشه بندی EM و بررسی صحت آن توسط نرم افزارداده کاوی WEKA

منیژه جلالی[۱] و ابوالقاسم بزرگ نیا[۲]

چکیده

امروزه با گسترش حجم عظیم اطلاعات و آشنایی متخصصین با داده کاوی، نیاز روزافزون به تکنیک های این دانش در برخورد با اطلاعات مختلف را نمایان ساخته است . یکی از این تکنیک ها خوشه بندی و ارائه الگوریتم های کارا برای این تحلیل می باشد . دراین مقاله سعی شده است به طور مختصر به فرآیند الگوریم EM در خوشه بندی وبررسی صحت آن توسط نرم افزار تخصصی داده کاوی Weka پرداخته شود. نتایج بدست آمده از تحقیق به ما کمک می کند تا در خوشه بندی داده ها، به خصوص هنگامی که با داده های گمشده سروکارداریم ، مناسب ترین خوشه را برگزینیم.

واژه های کلیدی : داده کاوی، خوشه بندی، الگوریتم EM، داده گمشده .

۱– مقدمه

داده کاوی فرایند کشف روابط ناشناخته والگوهای مناسب در داده ها است و به عنوان یک روش بسیار کارا برای کشف اطلاعات از داده ها شناخته شده است. یکی از قدیمی ترین ریشه های داده کاوی علم آمار است. بطوری که می توان گفت اگر تخمین وپیش بینی جزء وظایف داده کاوی درنظر گرفته شوند تحلیل های آماری، داده کاوی رابیش از یک قرن اجرا کرده است و به عقیده بعضی داده کاوی ابتدا از آمار وتحلیل های آماری شروع شد. درعمل دو هدف اصلی داده کاوی شامل پیش بینی وتوصیف می باشد که خوشه بندی یکی از اهداف توصیف است.

یکی ازانواع الگوریتم های کارا درخوشه بندی، در حالتی تعداد خوشه های انتخاب شده تصادفی است ، الگوریتم EM می باشد. الگوریم EM، در اواخر سال های ۱۹۷۰ توسط روبین، دمپستر ولارد[۳]، معرفی و گسترش داده شد که روشی محاسباتی برای برآورد داده به خصوص داده های گمشده می باشد. ثابت شده است که این روش یک روش محاسباتی بسیار کارا است[۱ ,۵,٤, ٦ ,۷]. در بخش ۲ تعاریف مختصری از خوشه بندی و الگوریتم EM آورده شده است. در بخش ۳ سعی شده است پارامترهای این الگوریتم، جهت پیاده سازی، معرفی شود. و در انتها این الگوریم بر روی یک سری داده آزمایش شده است.

۲– خوشه بندی

خوشه بندی داده ها، شاخه ای از داده کاوی می باشد که شامل تکنیک هایی برای پیدا کردن گروه هایی کوچک از میان پایگاه های بزرگ داده ها می باشد. هدف از خوشه بندی مشخص کردن ساختار داده هایی است که طبقه بندی نشده اند. برای رسیدن به این هدف سعی می شود داده ها در گروه هایی دسته بندی شوند بطوری که تفاوت داده های درون یک

[۱]دانشجوی کارشناسی ارشد آمار ، دانشگاه آزاد اسلامی مشهد
[۲]گروه آمار،دانشکده علوم دانشگاه آزاد اسلامی مشهد

گروه مینیمم شده و تفاوت داده هایی که در گروه های مختلف وجود دارند ماکسیمم شود.

یکی از گام های اصلی برای خوشه بندی مجموعه ای از داده ها انتخاب الگوریتم و پیاده سازی آن می باشد. دربرخی الگوریتم ها تعداد خوشه ها از ابتدا باید تعیین شود. در این حالت بهتر است که مسئله چند بار با تعداد متفاوتی خوشه اجرا شود و نتایج باهم مقایسه شوند. روش هایی نیز برای تخمین تعداد مناسب خوشه ها وجود دارد. اما در بعضی الگوریتم ها، تعداد خوشه ها در خلال اجرا و برحسب توزیع داده ها تعیین می شود، یا تعداد اولیه ای از خوشه ها در ابتدا تعیین می شود و الگوریتم ممکن است آن راتغییر دهد.

یکی از انواع الگوریتم های خوشه بندی، خوشه بندی تفکیکی می باشد. برای الگوریتم های تفکیکی ما از چارچوب کلی ارائه شده ی EM[1] استفاده می کنیم. این الگوریتم بصورت گسترده، روشی قابل استفاده برای محاسبه برآوردهای ماکزیمم درستنمایی در محاسبات مکرر است.

الگوریتم EM با روش های ویژه و تخصصی برآورد داده های گمشده، بطور دقیق ارتباط داده شده است که در این روش ها، پارامترها مجدداً برآورد می شوند و این فرآیند تا آن جا ادامه پیدا می کند که به یک مقداری همگرا شود . انتخاب نام EM به این علت است که در هر تکرار الگوریتم یک مرحله امید ریاضی گیری و بعد ازآن یک ماکسیمم سازی انجام می گیرد.

۳– پارامترهای الگوریتم خوشه بندی EM

پارامترهای لازم برای این الگوریتم عبارتند از :

k: تعداد خوشه ها

y: مجموعه ای از داده های p بعدی که الگوریتم را روی آنها اجرا می کنیم $Y = \{y_1, y_2 ..., y_n\}$

p: بعد هر داده

E: تلورانس برای لگاریتم درستنمایی.

Maxiteration: ماکزیمم مقدار تکرار حلقه ی خارجی.

اگرچه ما تعداد خوشه ها را داریم، ولی این نمی تواند به عنوان تنها شرط پایان بکار برود، چون حتی بعد از رسیدن به این تعداد خوشه، این الگوریتم ها مرتباً خوشه ها را به منظور بهبود بخشیدنشان تغییر می دهند. پس ما نیاز به شرط توقف نیز داریم. برای این کار ما نیاز به دریافت یک تلورانس از کاربر هستیم که مشخص کننده ی این است که فاصله ی خوشه ها از همدیگر حداکثر چقدر می تواند باشد؟ اگر در هر مرحله این فاصله ی بین خوشه ها بیشتر از این مقدار بود، الگوریتم باید ادامه پیدا کند. ولی چون ممکن است که این حداکثر فاصله برای بعضی از داده ها قابل دستیابی نباشد، نیاز به گرفتن پارامتر دیگری نیز داریم که تضمین کند که الگوریتم حتماً متوقف می شود. رسیدن الگوریتم به هرکدام از دو شرط فوق به منزله ی توقف الگوریتم می باشد. اگر کاربر این دو پارامتر را مشخص نکند، بطور پیش فرض، بلافاصله بعد از رسیدن به k خوشه، الگوریتم متوقف می شود[۸,۲].

مدل احتمال به کاررفته دراین الگوریتم، توزیع نرمال است، زیرا فرض می کند که مجموعه ی داده ها، می توانند به عنوان یک ترکیب خطی از توزیع نرمال چند متغیره درآیند. تابع چگالی احتمال توزیع نرمال در یک بعد بصورت زیر است :

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

که این تابع برای داده های p بعدی بصورت زیر در می آید:

$$p(X) = \frac{1}{\sqrt{(2\pi)^p |\sum|}} e^{-\frac{1}{2}(X-\mu)^T \sum^{-1} (X-\mu)}.$$

آن را می توان بصورت ساده تر زیر نوشت:

$$P(X) = \frac{1}{\sqrt{(2\pi)^p |\sum|}} e^{-\frac{1}{2}\delta^2},$$

که درآن δ را فاصله ماهالانوبیس گوییم و داریم:

$$\delta^2 = (X-\mu)^T \sum^{-1} (X-\mu).$$

و چون دراینجا داده های ما ترکیبی ازk خوشه از داده های p بعدی هستند، در نتیجه تابع احتمال مدل ترکیبی نرمال، بصورت زیر خواهد بود:

[1] Expectation Maximization

$$P(X) = \sum_{i=1}^{k} W_i P(X \mid i),$$

که در آن :

$P(X \mid i)$ توزیع نرمال برای هرخوشه است و W_I وزنی از داده های کلی است که این خوشه آن را معرفی می کند و $(\sum_{i=1}^{k} W_I = 1)$.

خروجی هایی که این الگوریتم به ما می دهد عبارتند از :

$c[p \times k]$: ماتریس میانه های خوشه ها (k خوشه p بعدی).

$R[P \times P]$: کوواریانس هرکدام از ابعاد. (کوواریانس ها بین خوشه ها مشترک است) این ماتریس قطری است.

$W[k \times 1]$: وزن های خوشه ها است.

$X[n,1]$: ماتریس احتمال عضویت هرداده در خوشه هااست.

سوالی که در اینجا مطرح است این است که آیا می توان ازسایر توزیع های آماری نیز برای بهبود نتایج و کلی کردن الگوریتم ها استفاده کرد یا نه؟ جواب این سوال مثبت می باشد. برای این منظور می توان از توزیع های پیوسته ی هم خانواده ی توزیع نرمال استفاده کرد. هر کدام از این توزیع ها درشرایطی بهتر از دیگری عمل می کند. که این شرایط اغلب به تعداد داده هایی که توزیع روی آنها اعمال می شود، بستگی دارد. پس این عامل هم می تواند به عنوان عامل خوبی برای کلی کردن الگوریتم، در نظر گرفته شود.

۴- الگوریتمEM

حال که شناختی از پارامترهای مختلف EM بدست آوردیم، به بیان الگوریتم می پردازیم:

E STEP
C'=0, R'=0 ,W'=0 ,IIh=0
For i=1 to n
Sumpi=0
For j=1 to k

$$\delta_{ij} = (y_i - c_j)^T R^{-1} (y_i - c_j)$$

$$P_{ij} = [W_i / ((2\pi)^{p/2} \mid R \mid^{1/2})] \exp[-1/2\delta_{ij}]$$

$$sumpi = sumpi + p_{ij}$$

End for

$$X_i = p / sumpi$$

$$IIh = IIh + In(sumpi)$$

$$c' = c' + y_i X_i^T$$

$$W' = W' + X_I$$

End for
M step
For j=1 to k

$$C_J = C'_J / W'_J$$

For i=1 to n

$$R' = R' + (y_i - c_j) X_{ij} (y_i - c_j)^T$$

End for

$$R = R' / n$$

$$W = W' / n$$

الگوریتم محبوب ومشهور k-Mwans یک حالت خاص از EM است وقتی که R و W ثابت بوده و برابر باشند با :

W=1/K
R=I

۵- پیاده سازی وبررسی صحت الگوریتم EM

ابزارهای داده کاوی متعددی وجود دارد که می تواند ما را در رسیدن به هدف مورد نظر یاری دهد. در این مقاله نرم افزار تخصصی داده کاوی Weka را مورد استفاده قرار می دهیم.

نرم افزارweka درآدرس

http://www.cs.waikato.ac.nz/me/weka

در دسترس است .

بعد از نصب نرم افزار، مجموعه داده های IRIS که یکی از مجموعه داده های ارزیابی کننده معروف می باشد که از سایت http://archive.ics.uci.edu/ml/datasets.iris فراخوانی شده است، مورداستفاده قرار می دهیم. جدول زیر قسمتی از فایل داده های مذکور را نشان می دهد .

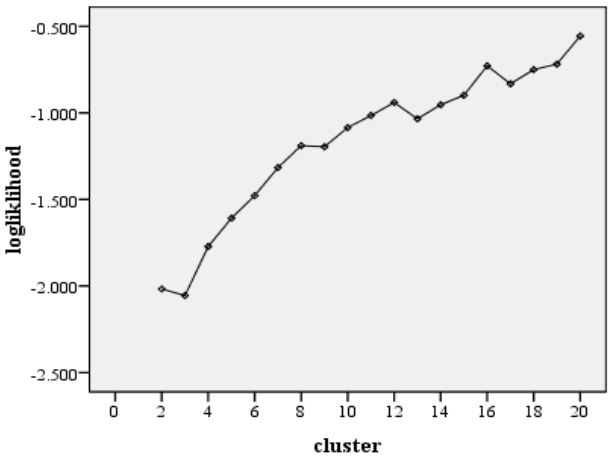
Viewer

Relation: iris

No.	sepallength Numeric	sepalwidth Numeric	petallength Numeric	petalwidth Numeric	class Nomina
1	5.1	3.5	1.4	0.2	Iris-se.
2	4.9	3.0	1.4	0.2	Iris-se.
3	4.7	3.2	1.3	0.2	Iris-se.
4	4.6	3.1	1.5	0.2	Iris-se.
5	5.0	3.6	1.4	0.2	Iris-se.
6	5.4	3.9	1.7	0.4	Iris-se.
7	4.6	3.4	1.4	0.3	Iris-se.
8	5.0	3.4	1.5	0.2	Iris-se.
9	4.4	2.9	1.4	0.2	Iris-se.
10	4.9	3.1	1.5	0.1	Iris-se.

شکل ۱ – بخشی از مشاهدات

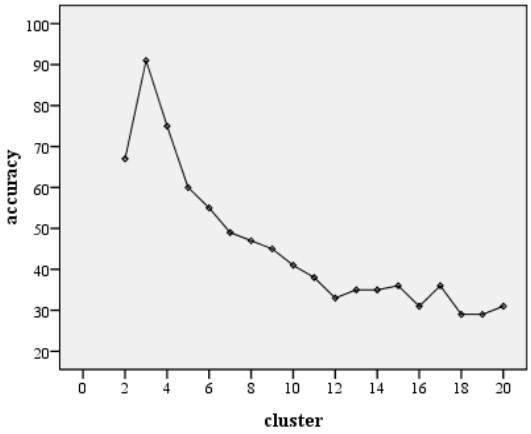
برای این مجموعه داده ، ۱۵۰ مشاهده[1] و ۵ صفت[2] وجود دارد. با ورود به پنجره cluster و درخواست الگوریتم EM و انتخاب خوشه های متفاوت، مقادیر لگاریتم درستنمایی(log-liklihood) را در خروجی خواهیم داشت. نتایج به صورت شکل ۲ زیر قابل مشاهده است.



شکل ۲ ـ نمودار رابطه بین تعداد خوشه ها و مقدار log- likelihood

همان طور که مشاهده می شود، با افزایش تعداد خوشه ها مقدار لگاریتم درستنمایی(log-liklihood) زیاد می شود و تنها در حالتی که تعداد خوشه را ۳ در نظر بگیریم در پایین ترین مقدار خود قرار دارد. یعنی حداکثر فاصله ی بین خوشه ها در حالتی اتفاق می افتد که تعدادخوشه ها را ۳ درنظر بگیریم. در واقع مقدار لگاریتم درستنمایی(log-liklihood) نشان دهنده ی تغییرات درون خوشه ها می باشد.

می توانیم جهت بالا بردن صحت کار، این نتیجه را با ستون آخر داده ها که به نام class نام گذاری شده است، و در آن یک خوشه بندی فرضی انجام شده بررسی کنیم. برای این منظور، نتایج مربوط به درصد صحت الگوریتم که در انتهای خروجی قابل مشاهده است را برای هر تعداد خوشه مورد نظر در نموداری به شکل زیر رسم می کنیم.



شکل ۳ ـ نمودار رابطه بین تعداد خوشه ها و صحت الگوریتم

همانطور که از روی نمودار مشخص است، در حالتی که تعداد خوشه ها را ۳ درنظر بگیریم، صحت الگوریتم در حدود ۹۱٪ خواهد بود. این می رساند که الگوریتم EM توانسته با درصد بالایی خوشه بندی مورد نظر را به درستی انجام دهد، و توانسته ۹۱٪ از داده ها را به درستی در خوشه های مربوطه جای دهد.

حال اگر تعدای از مشاهدات را گمشده درنظر بگیریم، همان نتیجه ای را خواهیم گرفت که از الگوریتم EM انتظار داشتیم. هیچ گونه تغییری در فرایند خوشه بندی ایجاد نخواهد شد. در این جا نیز انتخاب ۳ خوشه بالاترین صحت را به ما خواهد داد و این از مزیت های الگوریتم EM می باشد که به آن اشاره کردیم.

[1] Instances
[2] Attribute

۶- نتیجه گیری

در حالتی که تعداد نمونه زیاد و یا داده گمشده و ناقص وجود داشته باشد، استفاده از الگوریتم EM ما را به سمت انتخاب بهترین خوشه هدایت می کند. این الگوریتم همگرایی قابل اعتمادی دارد. یعنی با شروع از هر تعداد خوشه، همگرایی تقریباً همیشه به یک نقطه ماکسیمم موضعی حتمی است. مگر اینکه با بد شانسی تعداد خوشه ها را انتخاب کرده باشیم. همچنین در این الگوریتم می توان توابع آماری مختلفی را بر روی مشاهدات امتحان و بهترین تابع را انتخاب نمود.

منابع

[۱]. احمدی، جعفر(۱۳۸۲). طراحی یک زبان تولید پرس وجو انعطاف پذیر برای داده کاوی. کارشناسی ارشد. دانشگاه تهران.

[۲]. شیخائی، محمدصادق (۱۳۸٦). خوشه بندی سری های زمانی با استفاده از الگوریتم ژنتیک. کارشناسی ارشد، دانشگاه صنعتی امیر کبیر.

[3]. Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. Journal of The Royal Statistical Society, 39(1), 1-38.

[4]. Fayyad, U., Shapiro, G. P. and Smyth, P. (1996). Knowledge discovery and data mining: Towards a unifying framework. Proceedings of the second International conference on knowledge Discovery and data mining. Portland, Oregon, August, 2-4.

[5]. Hand. D.J. (1998). Review of data mining. The American statistician, 52, 112-118.

[6]. Heikki Mannila. (1997). Methods and problems in data minig.

[7]. Norwati M.and Jalali, M. (2009). Navigation Patterns Mining Approach based on Expectation Maximization Algorithm.

[8]. Ordonez, C. and Cereghini, P. (2000). SQLEM: Fast Clustering in SQL using the EM Algorithm.