Automatic Clustering of Mixed Data Using Genetic Algorithm

M.Yaghini* & M.Vard

Masoud Yaghini, Assistance professor of School of Railway Engineering - Iran University of Science and Technology
Mahdi Vard, MSc, School of Railway Engineering - Iran University of Science and Technology

Keywords

Data mining
Clustering
Mixed data
Genetic algorithm
Davies-Bouldin index

ABSTRACT

In the real world clustering problems, it is often encountered to perform cluster analysis on data sets with mixed numeric and categorical values. However, most existing clustering algorithms are only efficient for the numeric data rather than the mixed data set. In addition, traditional methods, for example, the K-means algorithm, usually ask the user to provide the number of clusters. In this paper, we propose a new method to cluster mixed data and automatically evolve the number of clusters as well as clustering of data set. In the proposed method, Davies-Bouldin Index is used as fitness function and we use the genetic algorithm to optimize fitness function. Also, we use a more accurate distance measure for calculating the distance between categorical values. The performance of this algorithm has been studied on real world and simulated data sets. Comparisons with other clustering algorithms illustrate the effectiveness of this approach.

© 2012 IUST Publication, IJIEPM. Vol. 23, No. 2, All Rights Reserved

*Corresponding author. Masoud Yaghini
Email: yaghini@iust.ac.ir
مسعود یقینی و همی فرد

چکیده:

مطالعه خوشبندی که منظور کمینه کردن مجموع مجدر احراز، یک مساله غیر خالص و غیر مشخص تابع و دارای تعاریف زیادی فقط به هنگام محیط است. در مقابل خوشبندی در دنیای واقعی، اغلب با مجموعه داده‌های مواج احراز که از ترکیبی از مقادیر عددی و دستی متفاوت شکل می‌شود. در حالتی که اغلب روش‌های خوشبندی موجود نیاز برای داده‌های عددی از کارایی مناسب برخوردند و قابلیت استفاده بر روی داده‌های متنوع را ندارند. از سوی دیگر، پیش‌بینی روش‌های سنتی، جادوی خوش‌بینه ها به عنوان ورودی از کاربر طلب می‌کند. در حالتی که پیش‌بینی مورد تعداد خوش‌بینه برای کاربر مقادیر نامل ممکن است و حذف‌زدین مقدار این نیز به خصوص در مورد موجودیت بزرگ کاری مشکل و حتی غیرممکن می‌شود. در این مقاله قصد داریم تا به‌صورتی از روش واقعی جهت استفاده در آن کاربردی گیری استاندارد دستیایی، روش‌های خوشبندی از داده‌های متنوع ارائه نماییم که نزدیک به تعریف داده‌های خوش‌بینه با معنا ورودی کمک می‌شود. در این مقاله روش‌بندی را به عنوان مقدار ساخته‌ای در نظر گرفته‌شده و به Davies-Bouldin مشاهده شده و به منظور جستجوی فضای جواب از کویت زنیک استفاده می‌شود. برای ارزیابی عملکرد کویت از دو گروه از داده‌های استاندارد و شیپ‌سازی شده استفاده شده است. نتایج بدست آمده، عملکرد بسیار بالایی کویت می‌شاند که نشان می‌دهد

1. مقدمه

خوشبندی عبارت از دستگاهی است که برای خوش‌بینه در اشیاء به منظور ایجاد گروه‌هایی از اشیاء تحت عنوان خوش‌بینه است. به نظر می‌رسد که این کار شبیه به همانندی در خوش‌بینه متغیر گزارشی می‌باشد. نتایج این اشخاص مختلفی از یکدیگر می‌باشد. روش‌بندی خوش‌بینه با توجه به روش‌بندی یکی از این استفاده می‌شود، به اینکه متغیر تغییر نشان می‌دهد که از آن می‌توان به روش‌بندی با افراد داده‌ها گروه سلسله را

پایگاه دیجیتال مقاله: دکتر جواد یقینی، استادیار، دانشکده مهندسی

yaghini@iust.ac.ir

فهرست نویسندگان: مهدی زر، کارشناس ارشد دانشکده مهندسی راه آهن، دانشگاه علم و صنعت ایران Mahdizavad@yahoo.com

۱ Hierarchical Methods
۲ Grid-based Methods
۳ Density-based Methods
۴ Model-based Methods
۵ Categorical Methods

123-شماره ۲۲-۱۳۹۱-بهمن ماه

نشریه بین المللی مهندسی صنایع و مدیریت تولید

شماره ۲۲ - جلد ۲۳ - بهمن ۱۳۹۱

http://IJIEPM.iust.ac.ir/

ISSN: 2008-4870

Davies-Bouldin

https://ijiepm.iust.ac.ir

مهدی زر، کارشناس ارشد دانشکده مهندسی راه آهن، دانشگاه علم و صنعت ایران

Mahdizavad@yahoo.com
روی داده‌های مختلط را داده‌ای بسیار مورد توجه و حائز اهمیت خود داره، بنابراین در حال حاضر به‌صورت عمده بر اینهای توپولوژی، سه مورد استفاده از تکنیک‌های زنگی که در تولید اینجا مورد استفاده قرار گرفته‌اند، روندهای مختلفی دارا داره. در اینجا باید به چنین نوشت: 

1) فست جینتیک ک-میانگین (Fast Genetic K-means Algorithm)
2) اینکrementال جینتیک ک-میانگین (Incremental Genetic K-means Algorithm)
3) الگوریتم زنگی کامپیوتری (Algorithm Genetic K-means)
روش‌های خوشه‌بندی داده‌های مخلوط

روش‌های خوشه‌بندی داده‌های مخلوط به چندین روش مختلفی تقسیم می‌شوند که عبارتند از:

1. **روش‌های خوشه‌بندی مبتنی بر افزایش داده‌ای**
   - **Genetic Weighted K-means Algorithm (GWKMA)**
   - **Improved Genetic Algorithm (IGA)**
   - **Similarity Based Agglomerative Clustering (SBAC)**

2. **روش‌های خوشه‌بندی داده‌های مخلوط**
   - **Simulated Annealing**
   - **Genetic Weighted K-means Algorithm**
   - **Improved Genetic Algorithm (IGA)**
   - **Similarity Based Agglomerative Clustering (SBAC)**

3. **روش‌های خوشه‌بندی بیش‌پهلوی**
   - **Squeezed Algorithm**

4. **روش‌های خوشه‌بندی کلیه**
   - **Genetic Weighted K-means Algorithm**
   - **Improved Genetic Algorithm (IGA)**
   - **Similarity Based Agglomerative Clustering (SBAC)**

**آدرس:** https://ijiepm.iust.ac.ir/article-1-82644-f.html

---

**منبع:** بهرامی، م. (1395). "روش‌های مختلف خوشه‌بندی داده‌های مخلوط". مجله‌ی علمی تخصصی بهره‌برداری از داده‌های مخلوط، شماره‌ی ۲، صفحات ۵۲-۶۴.
محاسبه مراکز خوشه‌ها برای داده‌های مختلط

در روش خوشه‌سازی با مختصات خوشه‌ها در این مقایسه ارائه شده با توجه به تعمیر مراکز خوشه‌ها، مقدار مرکزی به‌دست آمده درهای مشاهده‌ای مقدار مشاهده‌ای و مقدار دسته‌ی استفاده‌ای از نظر تطبیقی مستقل است. این مقدار دسته‌ی استفاده‌ای براساس توزیع کلی آنها در سرتاسر جمعه‌داهای تغییر می‌شود، باقی مقدار قابل‌توجهی از تغییر می‌شود. یک مقادیر متغیر خوشه‌ها به دست می‌آید.

در روش خوشه‌سازی با مختصات خوشه‌ها در این مقایسه ارائه شده با توجه به تعمیر مراکز خوشه‌ها، مقدار مرکزی به‌دست آمده درهای مشاهده‌ای مقدار مشاهده‌ای و مقدار دسته‌ی استفاده‌ای از نظر تطبیقی مستقل است. این مقدار دسته‌ی استفاده‌ای براساس توزیع کلی آنها در سرتاسر جمعه‌داهای تغییر می‌شود، باقی مقدار قابل‌توجهی از تغییر می‌شود. یک مقادیر متغیر خوشه‌ها به دست می‌آید.

در روش خوشه‌سازی با مختصات خوشه‌ها در این مقایسه ارائه شده با توجه به تعمیر مراکز خوشه‌ها، مقدار مرکزی به‌دست آمده درهای مشاهده‌ای مقدار مشاهده‌ای و مقدار دسته‌ی استفاده‌ای از نظر تطبیقی مستقل است. این مقدار دسته‌ی استفاده‌ای براساس توزیع کلی آنها در سرتاسر جمعه‌داهای تغییر می‌شود، باقی مقدار قابل‌توجهی از تغییر می‌شود. یک مقادیر متغیر خوشه‌ها به دست می‌آید.

در روش خوشه‌سازی با مختصات خوشه‌ها در این مقایسه ارائه شده با توجه به تعمیر مراکز خوشه‌ها، مقدار مرکزی به‌دست آمده درهای مشاهده‌ای مقدار مشاهده‌ای و مقدار دسته‌ی استفاده‌ای از نظر تطبیقی مستقل است. این مقدار دسته‌ی استفاده‌ای براساس توزیع کلی آنها در سرتاسر جمعه‌داهای تغییر می‌شود، باقی مقدار قابل‌توجهی از تغییر می‌شود. یک مقادیر متغیر خوشه‌ها به دست می‌آید.
 Auswahl مقادیر عم‌ل‌کرده‌های تغییرتکی در روش پیش‌نهایتی از مجموعه به‌کارگرفته شده در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva سطح در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactiva در مقداری عم‌ل‌کرده‌های درactive
جدول 1. مقدار پارامترهای ورودی الگوریتم پیشنهادی

<table>
<thead>
<tr>
<th>پارامتر ورودی</th>
<th>مقدار پارامتر</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>$k_{min}$</td>
</tr>
<tr>
<td>15</td>
<td>$k_{max}$</td>
</tr>
<tr>
<td>25</td>
<td>$\text{max}(\text{gen})$</td>
</tr>
<tr>
<td>0.5</td>
<td>$\psi_C$</td>
</tr>
<tr>
<td>0.01</td>
<td>$\psi_D$</td>
</tr>
<tr>
<td>40</td>
<td>اندازه جمعیت</td>
</tr>
</tbody>
</table>

$S_i = \frac{1}{|C_i|} \sum_{x \in C_i} ||x - z||_2$  \hspace{1cm} (8)

$R_i = \text{Max} \left[ \frac{S_i + S_j}{d_{ij}} \right]$  \hspace{1cm} (9)

$d_{ij} = d(C_i, C_j) = ||z_i - z_j||_2$  \hspace{1cm} (10)

$DB_r = \frac{1}{k_r} \sum_{i=1}^{k_r} R_i$  \hspace{1cm} (11)

$Fitness(Ch_i) = \frac{1}{DB_r}$  \hspace{1cm} (12)

که در روابط فوق $C_i$، $C_j$ نمایانگر خونه های $i$ و $j$ هستند، $d_{ij}$ فاصله مرکز خونه های $i$ و $j$ و $DB_r$ مقدار شاخص DB برای کروموزوم $r$ است.

4. آزمایش الگوریتم پیشنهادی

در این بخش نتایج اجرای الگوریتم خونه برده به الگوریتمهای با روش مقدار داده های استاندارد و در داده های شیب سازی شده با روش مقدار داده های استاندارد در داده های شیب سازی شده در ادامه می‌باشد، روش پیشنهادی نسبت به روش‌های قبلی بهترین راههای برازندگی است.

1- تنظیم پارامترهای ورودی الگوریتم با استفاده از مدل شیب سازی، صورت گرفته است. مقدار پارامترهای ورودی الگوریتم شامل مقدار جمعیت، $\psi_C$ و $\psi_D$، $\text{max}(\text{gen})$ و $k_{min}$ می‌باشد. این مقدارها مبتنی بر برازندگی می‌باشند. در جدول 1-2 نتایج حاصل از خونه برده این مجموعه داده را به وسیله روش خونه برده پیشنهادی در این مقاله نشان می‌دهد.

1-2-1 داده‌های بیماران قلبی

این داده‌ها اطلاعات مربوط به تعداد از بیماران قلبی را شامل می‌شود و در کلینیک‌های زیادی تولید شده است. پایگاه داده اصلی این داده‌ها مربوط به کلاس‌های متنوع در مورد خونه‌های باید برای کلاس‌های مختلفه برای بهبود کلاس‌های مختلف متنوع کلاس‌های مختلفه برای این مقادیر از $k_{min}$ و $k_{max}$ کمک می‌کند. در این مقاله نشان داده می‌شود که کلیه استفاده‌های این داده را به وسیله روش خونه برده پیشنهادی در این مقاله نشان می‌دهد.

شیب سازی [19] و [20]، روش‌های [21]، [22]، [23] و [24]، روش‌های [25] و [26]، روش‌های [27] و [28] است. مقدار $k_{min}$، $k_{max}$ و $\text{max}(\text{gen})$ در این مقاله نشان داده می‌شود که $k_{min}$ می‌باشد. در این مقاله نشان داده می‌شود که کلیه استفاده‌های این داده را به وسیله روش خونه برده پیشنهادی در این مقاله نشان می‌دهد.

1 http://archive.ics.uci.edu
2 Heart Disease Data
جدول 2. مقایسه نتایج روش‌های مختلف بر روی مجموعه داده کارتهای اعتباری

<table>
<thead>
<tr>
<th>دقت</th>
<th>تعداد داده هایی که در خوشه مورد انتخاب قرار گرفتند</th>
<th>نام الگوریتم</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.85</td>
<td>587</td>
<td>روش پیشنهادی</td>
</tr>
<tr>
<td>0.83</td>
<td>572</td>
<td>روش موده و اسیگنر</td>
</tr>
<tr>
<td>0.81</td>
<td>559</td>
<td>algCEBMC</td>
</tr>
</tbody>
</table>

بدین نسبت به دیگر روش‌ها بر روی داده های موسسه اعتباری سازی شده

در این بخش جهت سنجش نتایج الگوریتم پیشنهادی، مجموعه داده‌هایی از 2000 تا 12000 داده شده‌است.

روش خوشه‌بندی هاوانگ عبارت از k-prototypes است که به داده‌های مماثل می‌تواند مشخصات دسته‌ای و شکل‌های داده را مشخص کند. این الگوریتم داده‌های دارای 390 نمونه است که به دو کلاس تقسیم می‌شوند: کلاس منفی شامل 283 نمونه و کلاس مثبت شامل 307 نمونه.

نتایج حاصل از خوشه‌بندی این مجموعه داده نشان می‌دهد که الگوریتم algCEBMC و k-prototypes بهترین نتایج را در مجموعه داده های کارتهای اعتباری استفاده کرده است. در جدول 2 اورده شده است. مقادیر این نتایج نشان می‌دهد که گروه‌های خوشه‌بندی گروه‌های داده را مورد نظر قرار داده‌اند.

آماری صورت گیرد، از دو شاخص اصلی توجه باید به شاخص مجموع مطلق خطای SSE و شاخص مشاهده DB بپردازیم. SSE شاخص برای یافتن مفهوم جوابی از یک ارائه داده می‌باشد و DB شاخص برای یافتن مفهوم جوابی از یک ارائه داده می‌باشد. SSE شاخص برای یافتن مفهوم جوابی از یک ارائه داده می‌باشد و DB شاخص برای یافتن مفهوم جوابی از یک ارائه داده می‌باشد. SSE شاخص برای یافتن مفهوم جوابی از یک ارائه داده می‌باشد و DB شاخص برای یافتن مفهوم جوابی از یک ارائه داده می‌باشد. SSE شاخص برای یافتن مفهوم جوابی از یک ارائه D...
شکل 3 نسبت مقدار شاخص DB برای روش پیشنهادی به مقدار این شاخص برای روش k-prototypes به‌همه‌پایه‌تته: مجموعه داده‌های 12000 تایی

شکل 4 نسبت مقدار شاخص DB برای روش پیشنهادی به مقدار این شاخص برای روش k-prototypes به‌همه‌پایه‌تته: مجموعه داده‌های 12000 تایی

شکل 5 نسبت مقدار شاخص DB برای روش پیشنهادی به مقدار این شاخص برای روش k-prototypes به‌همه‌پایه‌تته: مجموعه داده‌های 12000 تایی


مراجع


