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

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
خوشبندی خوادگر داده‌های مختلط با استفاده از الگوریتم زنیتک

مسعود یاقینی و هماید ورد

چکیده:

مطالعه خوشه‌بندی به منظور کمیمی کردن مجموعه مجزا احراز، یک مطالعه غیر خاطر و غیر محض. بوده و دارای تعداد زیادی نقطه بهینه محلی است. در مسائل خوشه‌بندی در دنبال ایجاد، اغلب با مجموعه داده‌های موجودی که از ترکیبی از مقداری داده و دستایی تشکیل شده‌اند در حالیکه اغلب روش‌های خوشه‌بندی موجود نتایب روی داده‌های داده‌های از کارایی مناسب برخوردارند و قابلیت استفاده در روش‌های مختلف را ندارند. این سؤال در دیگر ورش خوشه‌بندی سنتی، تعداد خوشه‌ها را به عنوان ورودی از کاربر طلب می‌کند. در حالیکه در پیش‌گاه مورد تعداد خوشه‌بندی بهتر گزینه‌های تحلیلی است و یک دستگاه سطحی از روی دقت تحریک جهت انتخاب گزینه‌های قابلیت درست‌سازی، روش‌های متنوعی برای روش‌های خوشه‌بندی داشته‌اند که نمی‌توان به تغییر تعداد خوشه‌ها به عنوان ورودی الگوریتم نشانده و قادر است همزمان با خوشه‌بندی داده، همین‌طور برای مورد خوشه‌ها را به عنوان ورودی Davies-Bouldin مستند. برای ارزیابی عملکرد الگوریتم از منظور جستجوی فضای جواب از الگوریتم زنیتک استفاده می‌شود. برای ارزیابی عملکرد الگوریتم از دو گروه از داده‌ها استاندارد و شبیه‌سازی شده استفاده شده است. نتایج بدست آمده، عملکرد بسیار بالای الگوریتم داده‌های را نشان می‌دهد.

کلمات کلیدی

خوشه‌بندی

خوشه‌بندی

الگوریتم زنیتک

Davies-Bouldin

شناخت

1. مقدمه

خوشه‌بندی عبارت از دستگاهی تعدادی از اشیاء به منظور ایجاد گروه‌هایی از اشیاء تحت عنوان خوشه است. به ترتیب‌های آمیزه درک خوشه سیاره هم و شیب موجود در خوشه‌سازی مختلف مثلاً متمایز از یکدیگر یافتند. روش‌های مختلف خوشه‌بندی با توجه به روش‌هایی که برای گروه‌بندی داده‌ها از آن استفاده می‌کنند، به این‌گونه مختلف تقسیم می‌شوند که از آن میان می‌توان به روش‌های مناسب بر اساس این داده‌ها، روش‌های سلسله‌بار.

2. Partitioning Methods

مرتبط به روش‌های بیشتر، روش‌های مبنا بر تراکم و روشنایی مبنا بر مدل آن‌ها نمود [1، 2].

از میان روشهای شیب پیچیده که برای خوشه‌بندی داده‌ها طراحی و ارائه شده است، اغلب آن‌ها به دو دسته خوشه‌بندی نوع خاصی از داده‌ها (داده‌های محدود، که صرفه جویی از نوع دسته‌بندی باشد) طراحی شده‌اند و بر روی داده‌های مختلف که اندور مختلف مشخصه داده و دسته ای را شامل می‌شوند، قابلیت کاربرد ندارند. از سوی دیگر بسیاری از مسائل دریابی واقعی و اغلب پایگاه‌های داده‌ای خوشه‌بندی به منظور تحلیل آنها ایجاد شدند. داده هایی از نوع مختلط را در خود جای دادهند. بنابراین روشهای خوشه‌بندی که قابلیت کار بر

نویسنده‌منوی مقاله: دکتر مسعود یاقینی، استادیار، دانشکده مهندسی راه آهن دانشگاه علوم و صنعت ایران

yaghini@iust.ac.ir

مهدی ورد، کارشناسی ارشد دانشکده مهندسی راه آهن، دانشگاه علوم و صنعت

mahdivard@yahoo.com

یران

1 Hierarchical Methods

2 Grid-based Methods

3 Density-based Methods

4 Model-based Methods

5 Categorical Methods
روش‌های خوشه‌بندی مبنی بر الگوریتم زنتیک

1- روش‌های خوشه‌بندی مبتنی بر الگوریتم زنتیک
کریستی و متفری [4] روش خوشه‌بندی مبتنی بر الگوریتم زنتیک (GKA) ارائه کرده‌اند که می‌توان برای حل مسائل خوشه‌بندی برای تعداد خوشه‌ها. 

2- روش‌های خوشه‌بندی مبتنی بر الگوریتم زنتیک
کلینیک (FGKA) 1 را با الگوریتم GKA ابزار نرم‌افزارهایی که این فرآیند روش در سیستم‌های ویژه انجام می‌دهد. GKA یک الگوریتم سری‌پر عامل است از جمله اینکه روش یکی از جمله الگوریتم‌های ویژه، که به همراه الگوریتم GKA کار می‌کند.

3- روش‌های خوشه‌بندی مبتنی بر الگوریتم زنتیک
کلینیک (IGKA) 2 که از الگوریتم GKA با عملکرد استفاده می‌کند.

4- الگوریتم زنتیک
کلینیک (HGA) که از الگوریتم GKA با عملکرد استفاده می‌کند.

5- الگوریتم زنتیک
کلینیک (Davies-Bouldin) 3 که از الگوریتم GKA با عملکرد استفاده می‌کند.

روش‌های خوشه‌بندی مبتنی بر الگوریتم زنتیک

1 Fast Genetic K-means Algorithm
2 Incremental Genetic K-means Algorithm

روش‌های خوشه‌بندی مبتنی بر الگوریتم زنتیک

1 Fast Genetic K-means Algorithm
2 Incremental Genetic K-means Algorithm

روش‌های خوشه‌بندی مبتنی بر الگوریتم زنتیک

1 Fast Genetic K-means Algorithm
2 Incremental Genetic K-means Algorithm
کلیه متغیرهای مشخصات روش پیشنهادات.

**درک و بررسی روش‌های حاوی ترتیب بهینه‌سازی**

روش‌های حاوی ترتیب بهینه‌سازی به‌طور کلی برای حل سیستم‌های تاریکه‌گیری و چپ‌افزارهای با بیشترین مقدار مشخصه‌های مختلفی در داده‌ها استفاده می‌شود. این روش‌ها به ترتیب بهینه‌سازی می‌توانند باعث بهترین نتایج در سیستم‌های تاریکه‌گیری و چپ‌افزارهای با بیشترین مقدار مشخصه‌های مختلفی در داده‌ها باشند.

**نتیجه‌گیری‌های نهایی**

نتیجه‌گیری‌های نهایی از استفاده روش‌های حاوی ترتیب بهینه‌سازی بیشتر از روش‌های ساده‌تری بهترین نتایج را تهیه می‌کنند. این نتایج به عنوان شایعه‌ای در برخی از رشته‌های علمی و فنی مطرح شده است.
خوشه‌بندی دورکار داده‌های مختلط با استفاده از الگوریتم زنتیک

1- روش مورد استفاده برای محاسبه فاصله یک دو مقدار
در روش خوشه‌بندی به‌شیوه‌دای در این مقاله، از تابع فاصله‌ای ارائه شده توسط احمد و دی[18] ۱۸۳ جهت محاسبه فاصله‌ای مختص و نیز محاسبه متوسط فاصله‌ها استفاده شده است و جهت پیاده‌سازی الگوریتم، در روش پیشنهادی فاصله به‌شیوه‌دای استفاده می‌گردد. این تعریف فاصله بین دو مقدار $A_1$ و $A_2$ به صورت زیر تعریف می‌شود:

$$\delta^2(x, y) = P_1(w|x) + P_1( - w|y)$$

2- تعریف ۲
فاصله بین دو مقدار $x$ و $y$ به صورت زیر تعریف می‌شود:

$$\delta^2(x, y) = P_1(o|x) + P_1( - o|y) - 1$$

که به رابطه فوق، آن زیر مجموعه ای از مقدار $A_1$ می‌گردد.

3- تعریف ۳
که به این آن مقدار عبارت $P_1(o|x) + P_1( - o|y)$ که با رابطه فوق، آن زیر مجموعه ای از مقدار $A_1$ می‌گردد.

$$\delta(x, y) = \left(\frac{1}{m} - 1\right) \sum_{j=1}^{m} \delta^2(x, y)$$

1. Crisp Clustering
2. Co-occurrence

۲۳- شماره
نشره‌ی بین‌المللی مهندسی صنایع و مدیریت تولید، شماره ۲۳- جلد ۱۳۹۱- دی ۱۳۹۱
amiento انجام داده شده. تحمیل ویژه از منطقهی ایده‌آلی به سطح ویژهای دیگری از منطقه‌های خوشه‌ای مساهته می‌شود. برای این درصد از تحلیل‌های اخیر، تعداد نسبی از این است. یکی از دو کلاسیکی بین داده‌های همبستهстоя زیر تعریف

\[
\Omega(X) = \sum_{i=1}^{N_i} \delta(X, A_i) + \sum_{i=1}^{N_i} \delta(X, A_i) + \sum_{i=1}^{N_i} \delta(X, A_i)
\]

در نهایت فاصله کل میان یک داده و یک مکان خوشه برای مجموع داده‌های همبستهстоя زیر تعریف

\[
D(d_i, C_j) = \sum_{i=1}^{m} (d_i - C_j)^2 + \sum_{i=1}^{m} \Omega(d_i, C_j)^2
\]

پس از انجام پردازه‌های روشن شده، در روش پیشنهادی کروموزوم را از دارد. در مورد پردازش هوش‌مندبای‌های

\[
\Omega(X) = \sum_{i=1}^{N_i} \delta(X, A_i) + \sum_{i=1}^{N_i} \delta(X, A_i) + \sum_{i=1}^{N_i} \delta(X, A_i)
\]

نهایت فاصله کل میان یک داده و یک مکان خوشه برای مجموع داده‌های همبستهстоя زیر تعریف

\[
D(d_i, C_j) = \sum_{i=1}^{m} (d_i - C_j)^2 + \sum_{i=1}^{m} \Omega(d_i, C_j)^2
\]

پس از انجام پردازه‌های روشن شده، در روش پیشنهادی کروموزوم را از دارد. در مورد پردازش هوش‌مندبای‌های

\[
\Omega(X) = \sum_{i=1}^{N_i} \delta(X, A_i) + \sum_{i=1}^{N_i} \delta(X, A_i) + \sum_{i=1}^{N_i} \delta(X, A_i)
\]

نهایت فاصله کل میان یک داده و یک مکان خوشه برای مجموع داده‌های همبستهстоя زیر تعریف

\[
D(d_i, C_j) = \sum_{i=1}^{m} (d_i - C_j)^2 + \sum_{i=1}^{m} \Omega(d_i, C_j)^2
\]
جدول ۱. مقادیر پارامترهای ورودی الگوریتم پیشنهادی

<table>
<thead>
<tr>
<th>پارامتر ورودی</th>
<th>مقادیر پارامتر</th>
</tr>
</thead>
<tbody>
<tr>
<td>حداکثر تعداد خوشه‌ها</td>
<td>2</td>
</tr>
<tr>
<td>(k_{max})</td>
<td></td>
</tr>
<tr>
<td>حداکثر تعداد خوشه‌ها (max-gen)</td>
<td>15</td>
</tr>
<tr>
<td>لحظه زایمینه</td>
<td>25</td>
</tr>
<tr>
<td>نرخ پژوهیدن</td>
<td>0.5</td>
</tr>
<tr>
<td>نرخ توزیع</td>
<td>0.01</td>
</tr>
<tr>
<td>اندازه جمعیت</td>
<td>40</td>
</tr>
</tbody>
</table>

tablاب هدف است. معکوس این شاخص به عنوان مقدار نابع بر اثر تغییر شده است. در رابطه مربوط به محاسبه شاخص DB در ادامه ارائه می‌گردد.

\[ S_i = \frac{1}{|C_i|} \sum_{x \in C_i} \| x - z_i \| \]  \hspace{1cm} (8)

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

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

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

\[ \text{Fitness}(Ch_i) = \frac{1}{DB_r} \]  \hspace{1cm} (12)

که در روابط فوق، \( C_i \) نماینده خوشه‌که \( i \)ام، \( z_i \) نشان‌دهنده مرکز خوشه‌که \( i \)ام و \( DB_r \) نماینده مقدار شاخص DB برای کروموزوم \( r \) است.

#### ۴. آزمایش الگوریتم پیشنهادی

در این بخش نتایج اجرای الگوریتم خوشه‌بندی پیشنهادی بر روی مجموعه داده های استاندارد و داده های شیب سازی شده (آزمایش دوم) ارائه شده است. در ادامه دقت عملکرد روش پیشنهادی نسبت به روش‌های پیشین گردیده است.

#### ۱-۱ تنظیم پارامترهای ورودی الگوریتم

با استفاده از میانیت‌های شاید، مقدار پارامترهای ورودی الگوریتم شامل تعداد جمعیتی، نرخ زایمینه، نرخ جهش، حداکثر تعداد نسل‌ها و نسبت تغییر گردید که مقداری از این پارامترها در جدول ۱ آورده شده است. همچنین در الگوریتم پیشنهادی انتخاب و الگوهای مربوط به اتصال روش مناسب برای انتخاب مقدار می‌گردد.

#### ۲-۴ نتایج الگوریتم بر روی داده های استاندارد

در این بخش نتایج اجرای الگوریتم خوشه‌بندی پیشنهادی بر روی مجموعه داده‌های استاندارد ارائه شده است. به منظور مقایسه تغییر علائم الگوریتم آزمایش شده با روش‌های قبلی، از جدول مقایسه‌های استاندارد الگوریتم خود از این استفاده قدره ان تها امکان مقایسه نتایج وجود خواهد داشت. این داده‌ها با م�یون داده‌ها

---

2. Heart Disease Data

نحوه بین المللی مهندسی صنایع و مدیریت تولید: شریف‌بنی-۱۳۹۴/جلد ۲۳-شماره ۲
جدول ۲: مقایسه نتایج روش‌های مختلف بر روی مجموعه داده کارت‌های اعتباری

<table>
<thead>
<tr>
<th>نام الگوریتم</th>
<th>دقت داده هایی که در خوش مرد انرژی قرار گرفته‌اند</th>
</tr>
</thead>
<tbody>
<tr>
<td>روش پیشنهادی ما</td>
<td>0.85</td>
</tr>
<tr>
<td>SBAC</td>
<td>0.75</td>
</tr>
<tr>
<td>ECOWEB</td>
<td>0.74</td>
</tr>
<tr>
<td>COBWEB/3</td>
<td>0.81</td>
</tr>
<tr>
<td>Huang</td>
<td>0.81</td>
</tr>
<tr>
<td>algCEBMC</td>
<td>0.81</td>
</tr>
</tbody>
</table>

سایز شده
در این بخش، جهت نشان دادن الگوریتم پیشنهادی، مجموعه داده‌های با ابعاد ۲۰۰۰۰ و ۱۲۰۰۰۰ داده شده‌است. شده و فارید خوشنویسی در مورد آن‌ها انجام گرفت. همچنین با تغییر نهایی تعریف فاصله بین مقادیر دسته‌ای در k-prototypes، روش خوشنویسی هوآنگ بنی روشن و ساختار جدیدی ایجاد کرد. در این الگوریتم، نیز برای خوشنویسی داده‌های مورد استفاده قرار گرفته و نتایج حاصله استخراج گردید. به منظور مقایسه الگوریتم پیشنهادی و روش k-prototypes، به همراه دقت دیده‌شده، در هر مورد یک نمودار ساخته شد. این نمودارها نشان می‌دهد که اگر در نمونه‌ریزی فوق الگوریتم پیشنهادی از شاخ همانندی کردن باید عناوین از این الگوریتم بر روی روش پیشنهادی بر روی داده‌های مجموعه قلبی به دیگر روش‌ها بر روی داده‌های مجموعه اعتباری استفاده

شکل ۱: مقایسه میزان دقت روش خوشنویسی بندی پیشنهادی نسبت به دیگر روش‌ها بر روی داده‌های مجموعه اعتباری

نتایج حاصل از روش پیشنهادی در این مجموعه داده نیز توسط روش پیشنهادی و دو روش خوشنویسی بندی داده های مخلوط که برای آزمایش نتایج خود از مجموعه داده های کارت‌های اعتباری استفاده کرده‌اند. در جدول ۳ آورده شده است. مقادیر این جدول نیز برای دیگر روش از روش پیشنهادی را نسبت به روش ارائه شده توسط موده و استیلکر [۲۶] تانید می‌کند.

۲.۲ داده‌های کارت‌های اعتباری

Australian Credit Approval
شکل ۶ نسبت مقدار شاخص SSE برای روش پیشنهادی به مقدار این شاخص برای روش DB به‌همون‌یافته: مجموعه داده‌های ۲۰۰۰ تایی

اما به ارای تمامی مجموعه داده‌های شبیه سازی شده، مقدار شاخص SSE برای روش پیشنهادی مقدار کمتری را به خود اختصاص داده است.

شکل ۷ نسبت مقدار شاخص SSE برای روش پیشنهادی به مقدار این شاخص برای روش DB به‌همون‌یافته: مجموعه داده‌های ۲۰۰۰ تایی

شکل ۸ نسبت مقدار شاخص SSE برای روش پیشنهادی به مقدار این شاخص برای روش DB به‌همون‌یافته: مجموعه داده‌های ۲۰۰۰ تایی

شکل ۹ نسبت مقدار شاخص SSE برای روش پیشنهادی به مقدار این شاخص برای روش DB به‌همون‌یافته: مجموعه داده‌های ۲۰۰۰ تایی

شکل ۱۰ نسبت مقدار شاخص SSE برای روش پیشنهادی به مقدار این شاخص برای روش DB به‌همون‌یافته: مجموعه داده‌های ۲۰۰۰ تایی

شکل ۱۱ نسبت مقدار شاخص SSE برای روش پیشنهادی به مقدار این شاخص برای روش DB به‌همون‌یافته: مجموعه داده‌های ۲۰۰۰ تایی

شکل ۱۲ نسبت مقدار شاخص SSE برای روش پیشنهادی به مقدار این شاخص برای روش DB به‌همون‌یافته: مجموعه داده‌های ۲۰۰۰ تایی

شکل ۱۳ نسبت مقدار شاخص SSE برای روش پیشنهادی به مقدار این شاخص برای روش DB به‌همون‌یافته: مجموعه داده‌های ۲۰۰۰ تایی

شکل ۱۴ نسبت مقدار شاخص SSE برای روش پیشنهادی به مقدار این شاخص برای روش DB به‌همون‌یافته: مجموعه داده‌های ۲۰۰۰ تایی

شکل ۱۵ نسبت مقدار شاخص SSE برای روش پیشنهادی به مقدار این شاخص برای روش DB به‌همون‌یافته: مجموعه داده‌های ۲۰۰۰ تایی

شکل ۱۶ نسبت مقدار شاخص SSE برای روش پیشنهادی به مقدار این شاخص برای روش DB به‌همون‌یافته: مجموعه داده‌های ۲۰۰۰ تایی

شکل ۱۷ نسبت مقدار شاخص SSE برای روش پیشنهادی به مقدار این شاخص برای روش DB به‌همون‌یافته: مجموعه داده‌های ۲۰۰۰ تایی
در این مقاله یک روش خوشه‌بندی داده‌های مختلط مبنی بر الگوریتم زنتیک ارائه شده است. در روش پیشنهادی، برخی از اکثریت روش‌های خوشه‌بندی داده‌های مختلط که از تعداد فاصله صفر و یک برای اندازه‌گیری فاصله بین داده‌های بهره‌مندی، از تعریف دیقیتری جهت سنجه فاصله بین داده‌های دستیای و نیز محاسبه مراکز خوشه‌ها استفاده شده است. و سپس اجزای الگوریتم زنتیک (عملکردی تاپیکت و چهار) مناسب با ساختار جدید نامیش مراکز خوشه‌ها برای هر یک از داده‌های عددی و دستیای تعییر گردیده است.

۵ نتایج گیری

از دیگر مزایای روش پیشنهادی اینست که این روش به تعیین تعداد خوشه‌ها به‌عنوان ورودی الگوریتم نداده و قادر است با بهره‌گیری از قابلیت استخراج الگوریتم زنتیک در فضای جواب، ضمن خوشه‌بندی داده‌ها، مقدار بهینه تعداد خوشه‌ها را نیز محاسبه نماید که این ویژگی در مورد سبایی از یک دیتاباک واقعی که در آنها یک مجموعه داده‌های سایبری برای تعداد خوشه‌های معنی‌دار سرمایه‌دار در تاریخ‌های از استراتژی و توانایی الگوریتم پیشنهادی توانست داده‌های انتخاب و نیز داده‌های استراتژی‌های داده نشان از برتری یک روش و دقت بالاتر آن نسبت به سایر روش‌های خوشه‌بندی داده‌های مختلط دارد.

مراجع


