سیاوش حسن پور 921042006

:1

Timelines: (به موقع بودن) داده ها باید در یک چارچوب زمانی در دسترس باشد تا مفید واقع شوند.

Believability:(باور) ارزش داده ها باید در محدوده نتایج ممکن است تا مفید واقع شوند.

Interpretability: (قابلیت تفسیر) داده نباید خیلی پیچیده باشد که درک اطلاعات آن سخت شود.

دسترسی:داده باید در دسترس باشد.

:2

Ignoring the tuple : (نادیده گرفتن تاپل) این روش معمولا وقتی که برچسب کلاس از دست رفته است انجام می شود. این روش زیاد موثر نیست مگر این که تاپل شامل ویژگی های متعددی با مقدار از گم شده باشد.

Fill in the missing value manually : (دستی دادن مقادیردر ارزش از دست رفته) این رویکرد بسیار وقت گیر است و ممکن است یک کار درست برای این نوع داده نباشد، به ویژه هنگامی که ارزش از دست رفته به راحتی مشخص نباشد.

Using a global constant to fill in the missing value : ( با استفاده از یک ثابت جهانی برای مقداردهی مقدار از دست رفته) به جای تمام مقادیر ویژگی گمشده با استفاده از ثابت، مانند یک برچسب مانند «ناشناس» مقادیر گمشده را مثلا با «ناشناس» جایگزین می کنیم.

Using the attribute mean for quantitative (numeric) values or attribute mode for categorical (nominal) values, for all samples belonging to the same class as the given tuple : استفاده از ویژگی معنی برای مقادیر کمی (عددی) و یا حالت صفت برای مقادیر طبقه (اسمی) برای تمام نمونه های متعلق به همان کلاس تاپل داده

Using the most probable value to fill in the missing value : (با استفاده از مقدار محتمل ترین به مقدار از دست رفته) این مقدار ممکن است با رگرسیون تعیین شود .ابزارهای مبتنی بر استنتاج با استفاده از فرمالیسم بیزی، و یا درخت تصمیم گیری.

:3

a)

1)

13, 15, 16, 16,19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

2) Bin:

1: 13, 15, 16

2: 16, 19, 20

3: 20, 21, 22

4: 22, 25, 25

5: 25, 25, 30

6: 33, 33, 35

7: 35, 35, 35

8: 36, 40, 45

9: 46, 52, 70

1: 14.66

2: 18.33

3: 21

4: 24

5: 26.66

6: 33.66

7: 35

8: 40.33

9: 56

4) Bin:

1: 14.66, 14.66, 14.66

2: 18.33, 18.33, 18.33

3: 21, 21, 21

4: 24, 24, 24

5: 26.66, 26.66, 26.66

6: 33.66, 33.66, 33.66

7: 35, 35, 35

8: 40.33, 40.33, 40.33

9: 56, 56, 56

b)

Outliers در داده ها, ممکن است توسط خوشه شناسایی شده باشد، که در آن مقدارهای مشابه در یک گروه و یا "خوشه" قرارگرفته است . روش دیگر، ترکیبی از کامپیوتر و بازرسی انسان را می توان مورد استفاده قرار گیرد که در آن یک توزیع داده های از پیش تعیین شده اجرا شده است که اجازه می دهد تا کامپیوتر به شناسایی نقاط دورافتاده بپردازد.

c)

روش های دیگر که می توانند برای صاف کردن داده ها مورد استفاده قرارگیرند عبارتند از فرم های متناوب از مرج کردن. روش دیگر، سطل مساوی عرض که می توان برای پیاده سازی هر یک از اشکال مرج کردن که در آن محدوده فاصله از ارزش در هر بن ثابت است، استفاده کرد. روش های دیگر از مرج کردن شامل استفاده از تکنیک های رگرسیون برای صاف کردن داده ها است طبقه بندی تکنیک های کاتیون هم می تواند برای اجرای سلسله مراتب مفهومی که میتوان داده ها را نورد تا مفاهیم سطح پایین تر به مفاهیم سطح بالاتر صاف گردد.

:4

یکپارچه سازی داده ها شامل ترکیب داده ها از منابع مختلف به یک فروشگاه داده منسجم است

مسائل موجود در یکپارچه سازی :

Schema integration :(ادغام معماری) metadata از منابع داده مختلف باید به منظور مطابقت معادل اشخاص در دنیای واقعی یکپارچه شده است استفاده کند . این است که به عنوان مشکل نهاد شناسایی اشاره شده است.

Handling redundant data : (سیستم های انتقال داده های تکراری) ویژگی های مشتق شده ممکن است از کار برکنار شده، و نامگذاری ویژگی متناقض نیز ممکن است به موانعی در مجموعه داده ها و در نتیجه منجر شود. وجود داشتن تکراری در سطح تاپل ممکن است رخ دهد و در نتیجه نیاز به شناسایی و حل داشته باشد.

Detection and resolution of data value convicts : (تشخیص و تحلیل محکومین مقدار داده)