دانشگاه شهید بهشتی
دانشکده مهندسی برق و کامپیوتر

ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

پایان نامه کارشناسی ارشد مهندسی کامپیوتر
گرایش نرم‌افزار کامپیوتر

استاد راهنما :
دکتر اسلام ناظمی

استاد مشاور:
مهندس حسین علیزاده

ارائه دهنده :
مانی دوستدار

فهرست مطالب

- مقدمه و مفاهیم
- کارهای مرتبط
- روش پیشنهادی
  - ✓ کلیات
  - ✓ سازماندهی قوانین
  - ✓ عملکرد عامل‌ها
- شبیه‌سازی و ارزیابی
- نتیجه‌گیری و کارهای آینده

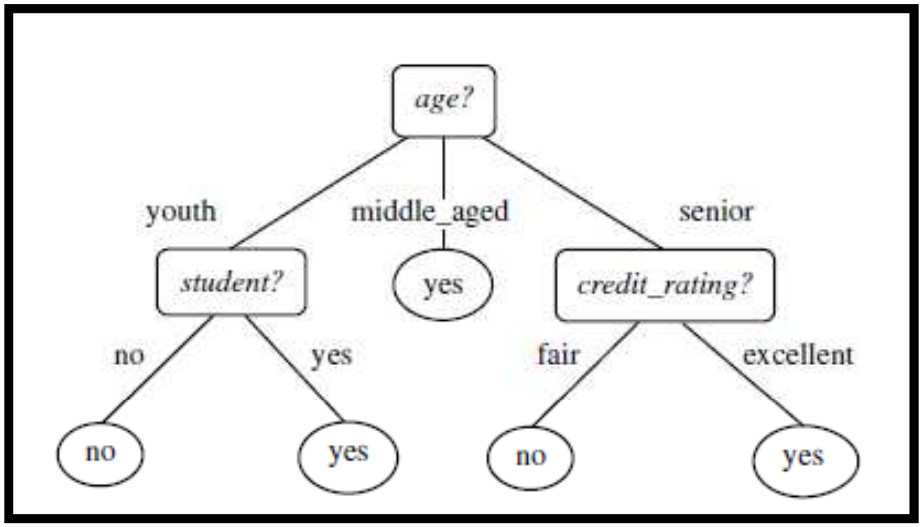
مقدمه

- امروزه سازمان-ها علاوه بر ذخیره-سازی داده-ها به اطلاعات سطح بالاتری احتیاج دارند تا بتوانند پیش-بینی-هایی انجام داده و تصمیماتی را اتخاذ کنند.
- داده‌کاوی: داده-کاوی نتیجه-تکامل طبیعی فن-آوری اطلاعات باشد و در واقع به استخراج دانش از حجم بالایی از داده اطلاق می-گردد.
- فعالیت-های داده-کاوی
  - ✓ فعالیتهای توصیفی داده‌کاوی
  - ✓ فعالیتهای پیشگویانه‌ی داده‌کاوی
    - ردەبندی داده‌ها

ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

رده‌بندی

- رده-بندی به فرآیند یافتن یک مدل(تابع) گفته می-شود که رده-های داده-ها و مفاهیم را توصیف و مشخص کند.
- مراحل رده‌بندی داده‌ها
  - ✓ یادگیری
  - ✓ آزمایش و استفاده
- نحوه نمایش مدل
  - ✓ مجموعه‌ای از قوانین
  - ✓ درخت تصمیم
  - ✓ فرمول ریاضی



ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

سیستم‌های مبتنی بر عامل

- سیستمی که در طراحی آن از مفهوم عامل استفاده شده است.
- عامل: بخشی از سیستم که به صورت خودکار وظایفی را که بر عهده دارد انجام می‌دهد.
- ویژگی‌ها:
  - ✓ محیط
  - ✓ ورودی و خروجی
  - ✓ سطح اختیار عامل
  - ✓ نحوه ارتباط عامل‌ها
    - ناظری بر فعالیت عامل وجود ندارد - داشتن هدف مشترک – ارسال پیام - اهمیت صحت اطلاعات دریافتی
    - سطوح مختلف برای عامل‌ها – تلاش برای رسیدن به یک هدف – مدیریت یک سطح از عامل‌ها توسط عامل دیگر

ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

طرح مسئله

- مشکل موجود
  - ✓ در درسترس نبودن مجموعه داده در یک محل و متفاوت بودن مجموعه داده‌ها از نظر برخی از ویژگی‌ها(تعداد صفات و تعداد نمونه‌ها)
- راه‌حل پیشنهادی
  - ✓ استفاده از سیستم‌های چند عاملی
  - ✓ استفاده از قالب مجموعه قوانین(تشکیل پایگاه جامع قوانین)

ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

ترکیب دو حوزه‌ی داده‌کاوی و سیستم‌های مبتنی بر عامل

- عامل‌های مبتنی بر داده‌کاوی
- داده‌کاوی مبتنی بر عامل(کاوش مبتنی بر عامل)
  - ✓ داده‌کاوی توزیع شده‌ی چند عاملی(استفاده از مدل همتا به همتا)

ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

ارزیابی و استفاده از قوانین

- استفاده از قوانین
  - ✓ استفاده از بهترین قانون
  - ✓ استفاده از K قانون با اولویت بالاتر
  - ✓ استفاده از همه‌ی قوانین
- مرتب‌سازی قوانین
  - ✓ روش‌های مبتنی بر پشتیبانی-صحت(CSA وACS)
  - ✓ روش‌های توزین قانون(WRA، LA و χ^2)
- انتخاب قوانین از مجموعه نتایج(حذف قوانین زائد)
  - ✓ قبل از تولید نتایج
  - ✓ بعد از تولید نتایج

ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

ترکیب روش‌های رده‌بندی

- در دسترس نبودن تمام حجم داده
  - ✓ ترکیب روش‌های رده‌بندی(یادگیری گروهی)
- حجم بالای مجموعه داده
  - ✓ بررسی بخش‌هایی از مجموعه داده
  - ✓ ترکیب نتایج
- کاهش پیچیدگی و افزایش کارایی(داده با ابعاد بالا)
  - ✓ تقسیم افقی و عمودی

ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

کلیات و ویژگی‌های اصلی

- ✓ استفاده از سیستم‌های مبتنی بر عامل برای تحلیل بهتر مسئله(نامتمرکز بودن داده‌ها و ناهمگونی آنها)
- ✓ بررسی بخشی از داده‌های حوزه مذکور به منظور تحلیل حوزه و مشخص کردن پارامترهای ارزیابی
- ✓ حضور انواع مختلف عامل‌ها به منظور انجام وظایف گوناگون در سیستم
- ✓ قالب نتایج به صورت مجموعه قوانین است(نتایج میانی و نهایی)
- ✓ اعتبارسنجی نتایج به صورت دوره‌ای

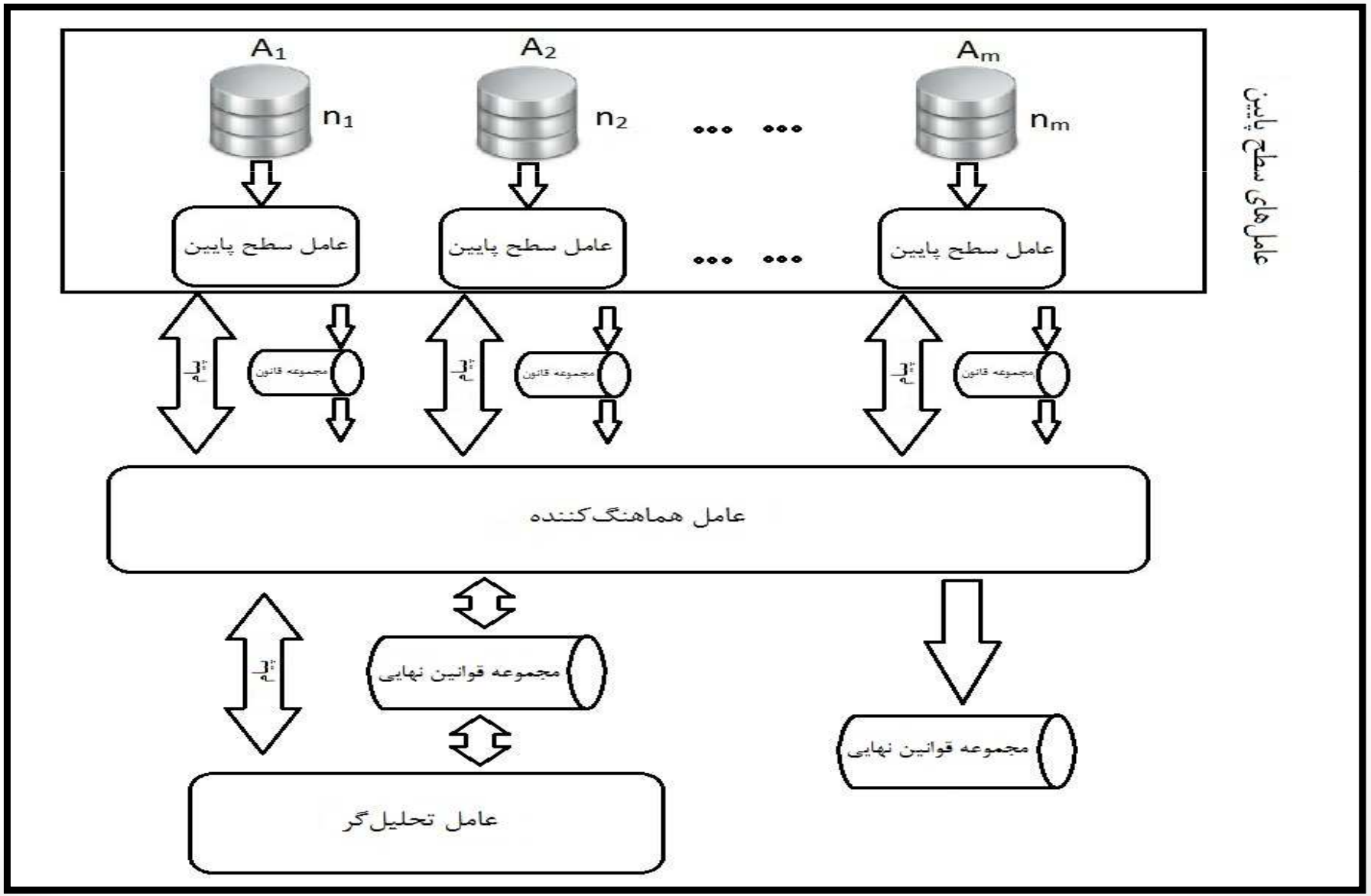
ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

عامل‌ها

- عامل‌های سطح پایین(Low level)
  - ✓ رده‌بندی اولیه‌ی داده‌ها و ارسال نتایج به عامل هماهنگ‌کننده
  - ✓ عدم ارتباط با عامل‌های دیگر در همان سطح
- عامل هماهنگ‌کننده(Coordinator)
  - ✓ دریافت نتایج از عامل‌های سطح پایین و تولید مجموعه قوانین نهایی
  - ✓ ارسال قوانین به عامل تحلیل‌گر برای اعتبارسنجی
- عامل تحلیل‌گر(Analyst)
  - ✓ دریافت نتایج از عامل هماهنگ‌کننده و اعتبارسنجی آن‌ها
  - ✓ حذف قوانین زائد

ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

نحوه ارتباط عامل‌ها با یکدیگر



ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

سازماندهی قوانین

- فاکتورهای مهم در رتبه‌بندی قوانین
- روش محاسبه‌ی امتیاز قوانین
- تعیین پارامترهای مهم برای ارزیابی قوانین

فاکتورهای مهم در رتبه‌بندی قوانین

- در سطح عامل سطح پایین
  - ✓ صحت(Confidence)
  - ✓ پشتیبانی(Support)
  - ✓ اندازه‌ی بخش پیشین(size of Antecedent)
- در سطح عامل هماهنگ‌کننده
  - ✓ تعداد نمونه‌ها
  - ✓ تعداد عامل‌های تولید کننده قانون

ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

روش مجاسبه‌ی امتیاز قوانین

- محاسبه‌ی امتیاز جزئی در عامل سطح پایین
  - ✓ صحت قانون(C)

$$C = \frac{n_{true}}{n_{support}}$$

  - ✓ پشتیبانی قانون(S)

$$S = \frac{n_{support}}{n_{total}}$$

  - ✓ نسبت تعداد صفت‌های بخش پیشین(A)

$$A = \frac{number\ of\ attributes_{rule}}{number\ of\ attributes_{total}}$$

  - ✓ محاسبه‌ی امتیاز جزئی

$$score_{partial} = w_1 C + w_2 S + w_3 A$$
$$0 < w_i \leq 1$$

ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

روش مجاسبه‌ی امتیاز قوانین(ادامه)

- اطلاعات قوانین
  - ✓ متن قانون(شرط و بدنه‌ی اصلی)
  - ✓ امتیاز جزئی محاسبه شده توسط عامل مربوطه
  - ✓ تعداد نمونه‌هایی که قانون از آنها تولید شده است
  - ✓ شماره‌ی عامل

روش محاسبه‌ی امتیاز قوانین(ادامه)

- بررسی وضعیت قانون جدید در مقایسه با مجموعه قوانین نهایی
  - ✓ شباهت و تناقض
  - ✓ حذف قوانین زائد
- محاسبه‌ی امتیاز کلی در عامل هماهنگ‌کننده
  - ✓ نسبت تعداد نمونه‌های عامل

$$0 < \frac{n_{agent}}{n_{max}} \leq 1$$

- ✓ محاسبه‌ی امتیاز کلی قانون

$$Score_{final} = \frac{\sum_{i=1}^{number\ of\ agents} n_i \times score_i}{number\ of\ agents \times n_{max}}$$

ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

تعیین پارامترهای مهم برای ارزیابی قوانین

- رتبه‌بندی قوانین
  - ✓ مشخص کردن ضرایب برای محاسبه‌ی امتیاز قانون توسط هر عامل سطح پایین

$$score_{partial} = w_1 C + w_2 S + w_3 A$$

$$0 < w_i \leq 1$$

- انتخاب قوانین
  - ✓ مشخص کردن مقدار آستانه برای فاکتورهای صحت و پشتیبانی به منظور انتخاب قوانین

تعیین پارامترهای مهم برای ارزیابی قوانین(رتبه‌بندی قوانین)

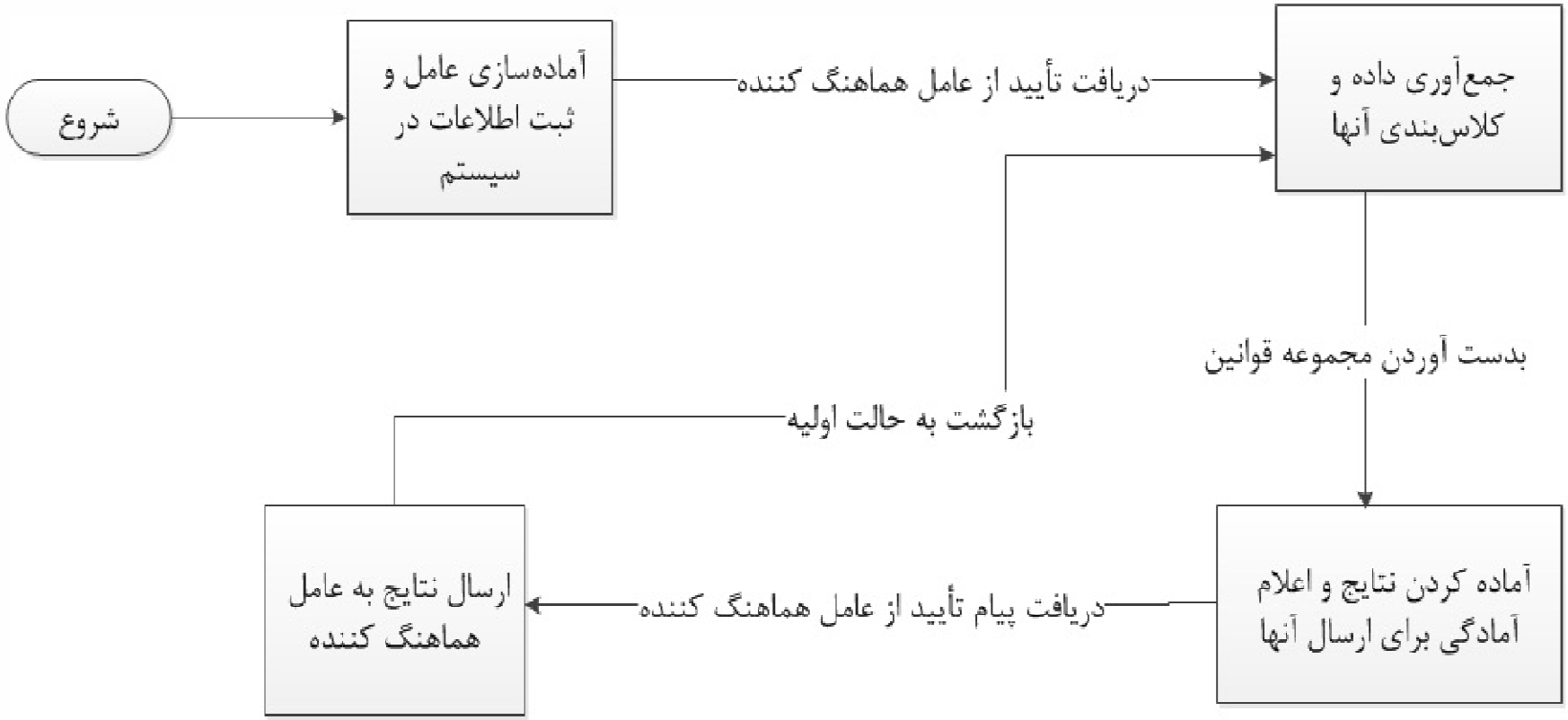
- طراحی آزمایش برای تعیین ضرایب به منظور رتبه‌بندی قوانین
  - ✓ انتخاب یک مجموعه داده برای تحلیل حوزه(یادگیری و آزمایش)
  - ✓ رده‌بندی مجموعه داده و بدست آوردن نتایج در قالب مجموعه قوانین
  - ✓ تعیین مقادیر مناسب برای ضرایب رابطه‌ی محاسبه‌کننده امتیاز قانون توسط عامل سطح پایین(W_iها)
  - ✓ رتبه بندی نتایج حاصل از رده‌بندی داده‌ها با استفاده از مقادیر مشخص شده برای ضرایب
  - ✓ آزمایش نتایج رتبه‌بندی شده با استفاده از مجموعه داده آزمایشی انتخاب شده
  - ✓ انتخاب بهترین ترکیب ضرایب

تعیین پارامترهای مهم برای ارزیابی قوانین(انتخاب قوانین)

- طراحی آزمایش برای تعیین مقادیر آستانه به منظور حذف قوانین زائد
  - ✓ تعیین مقادیر مناسب برای آستانه دو فاکتور صحت و پشتیبانی
  - ✓ انتخاب یک ترکیب از مقادیر آستانه و آزمایش مجموعه قوانین انتخاب شده با استفاده از مجموعه داده‌های آزمایشی در هر تکرار
  - ✓ انتخاب بهترین ترکیب برای مقادیر آستانه

شرح عملکرد عامل‌ها

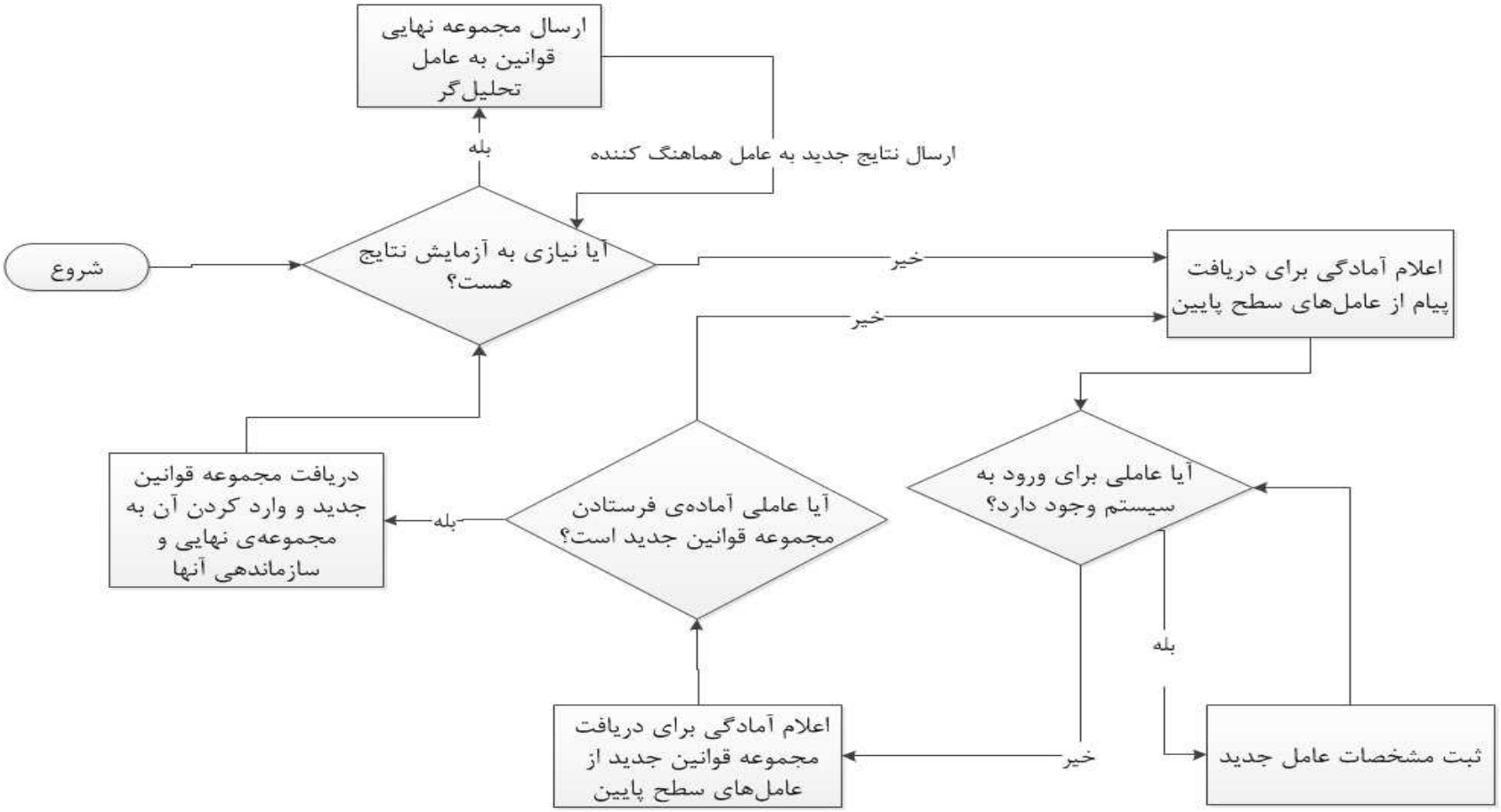
- عامل‌های سطح پایین



ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

شرح عملکرد عامل‌ها

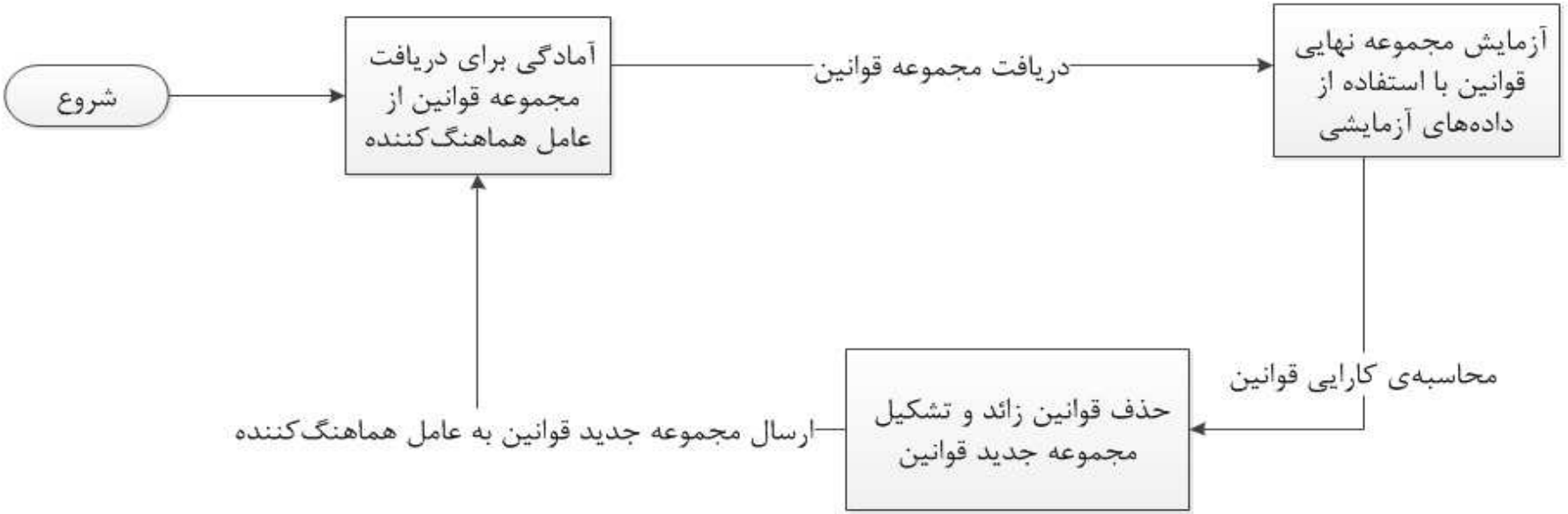
- عامل هماهنگ‌کننده



ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

شرح عملکرد عامل‌ها

- عامل تحلیل‌گر

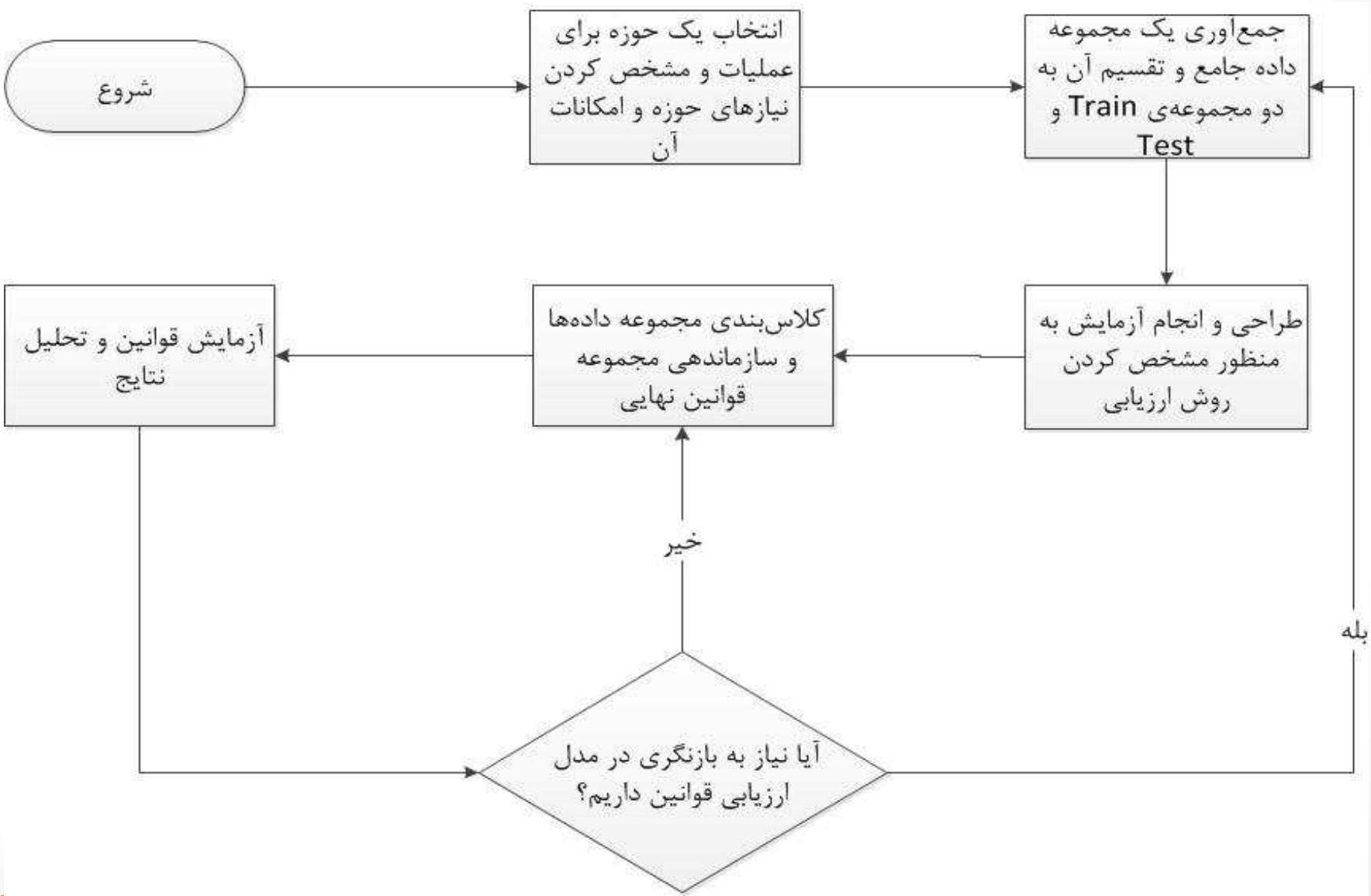


$$Efficiency = \frac{n_{correctly\ classified} - n_{incorrectly\ classified}}{n_{classified}}$$

$$-1 \leq Efficiency \leq 1$$

ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

نمای کلی روش پیشنهادی



ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

ابزارهای مورد استفاده

- مجموعه نرم‌افزاری WEKA
  - ✓ مجموعه‌ای از الگوریتم‌های یادگیری ماشین
  - ✓ ابزارهای پیش پردازش داده‌ها
  - ✓ نوشته شده به زبان جاوا

- بسته نرم‌افزاری Microsoft visual studio
  - ✓ پیاده‌سازی عامل‌ها
  - ✓ استفاده از زبان برنامه‌سازی C#.net

IKVM.NET

- بسته‌ی نرم‌افزاری IKVM
  - ✓ رابط بین عامل‌ها و نرم‌افزار WEKA
  - ✓ تبدیل کلاس‌های جاوا به فایل‌ها DLL

ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

شبیه‌سازی عامل‌ها

- برقراری ارتباط
  - ✓ عامل هماهنگ‌کننده دخالتی در وظایف عامل‌های سطح پایین ندارد
  - ✓ عامل‌ها تنها با ارسال پیام با یکدیگر ارتباط برقرار می‌کنند
- عامل سطح پایین
  - ✓ استفاده از روش PART برای رده‌بندی اولیه‌ی داده‌ها(دلیل: قالب نتایج الگوریتم)
  - ✓ تبدیل نتایج و اطلاعات تولید شده به یک قالب خاص برای پردازش آسان‌تر

روش ارزیابی

- مقایسه‌ی روش پیشنهادی با روش‌هایی که مجموعه داده را به صورت یکجا بررسی می‌کنند
- استفاده از یک مجموعه داده و تقسیم آن برای مراحل مختلف

✓ روش پیشنهادی

- مجموعه داده برای تحلیل حوزه(یادگیری و آزمایش)
- مجموعه داده برای انجام عملیات رده‌بندی و سازماندهی مجموعه نهایی قوانین(یادگیری و آزمایش)
  - ❖ تقسیم داده‌ها بین عامل‌های مختلف در تکرارهای متفاوت

✓ روش‌های مورد مقایسه

- مجموعه داده برای انجام عملیات رده‌بندی(یادگیری و آزمایش)

ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

مجموعه داده‌ی Adult

- مشخصات

تعداد صفت‌ها	نوع صفت‌ها	تعداد نمونه برای یادگیری	تعداد نمونه برای آزمایش	تعداد نمونه برای تحلیل حوزه
۱۲	مقدار قطعی	۳۲۵۶۰	۱۱۲۸۲	۵۰۰۰

- نتایج تحلیل حوزه

روش	CSA	ACS	New method
درستی روش	٪۷۶	٪۷۳	٪۸۰

✓ آستانه‌ی صحت:۰.۷

✓ آستانه‌ی پشتیبانی:۰.۰۲

$$score_{partial} = 0.2 \times C + S + 0.3 \times A$$

ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

مجموعه داده‌ی Adult

- درستی روش در پایان تکرارها

روش / تکرار	تکرار اول	تکرار دوم	تکرار سوم
درستی روش	۸۱.۴۰	۸۳.۲۹	۸۳.۱۹

- مقایسه‌ی نتایج

روش	روش پیشنهادی	روش PART	روش Random Forest	روش ZeroR	روش CR	روش Ridor	روش Decision Table	روش JRip	روش J48
درستی	۸۳.۱۹	۸۳.۱۴	۸۱.۹۷	۷۵.۶۴	۷۵.۶۴	۸۱.۶۵	۸۱.۷۸	۸۲.۵۳	۸۱.۹۱

ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

مجموعه داده‌ی Car

- مشخصات

تعداد صفت‌ها	نوع صفت‌ها	تعداد نمونه برای یادگیری	تعداد نمونه برای آزمایش	تعداد نمونه برای تحلیل حوزه
۷	مقدار قطعی	۹۶۳	۴۵۹	۳۰۶

- نتایج تحلیل حوزه

روش	CSA	ACS	New method
درستی روش	٪۸۳.۰۱	٪۸۳.۸۵	٪۸۳.۶۶

$$score_{partial} = 0.2 \times C + S + 0.5 \times A$$

ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

مجموعه داده‌ی Car

- درستی روش در پایان تکرارها

روش / تکرار	تکرار اول	تکرار دوم
صحت روش	۷۸	۸۵.۰۸

- مقایسه‌ی نتایج

روش	روش پیشنهادی	روش PART	روش Random Forest	روش ZeroR	روش CR	روش Ridor	روش Decision Table	روش JRip	روش J48
درستی	۸۵.۰۸	۸۵.۷۱	۶۰.۶۱	۶۷.۵۶	۶۷.۹۳	۷۹.۳۰	۷۶.۰۶	۶۴.۸۶	۷۶.۴۴

ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

مجموعه داده‌ی Nursery

- مشخصات

تعداد صفت‌ها	نوع صفت‌ها	تعداد نمونه برای یادگیری	تعداد نمونه برای آزمایش	تعداد نمونه برای تحلیل حوزه
۹	مقدار قطعی	۷۴۶۰	۳۵۰۰	۲۰۰۰

- نتایج تحلیل حوزه

روش	CSA	ACS	New method
درستی روش	٪۹۲.۴	٪۷۶.۸	٪۹۶

$$score_{partial} = 0.2 \times C + S + 0.5 \times A$$

ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

مجموعه داده‌ی Nursery

- درستی روش در پایان تکرارها

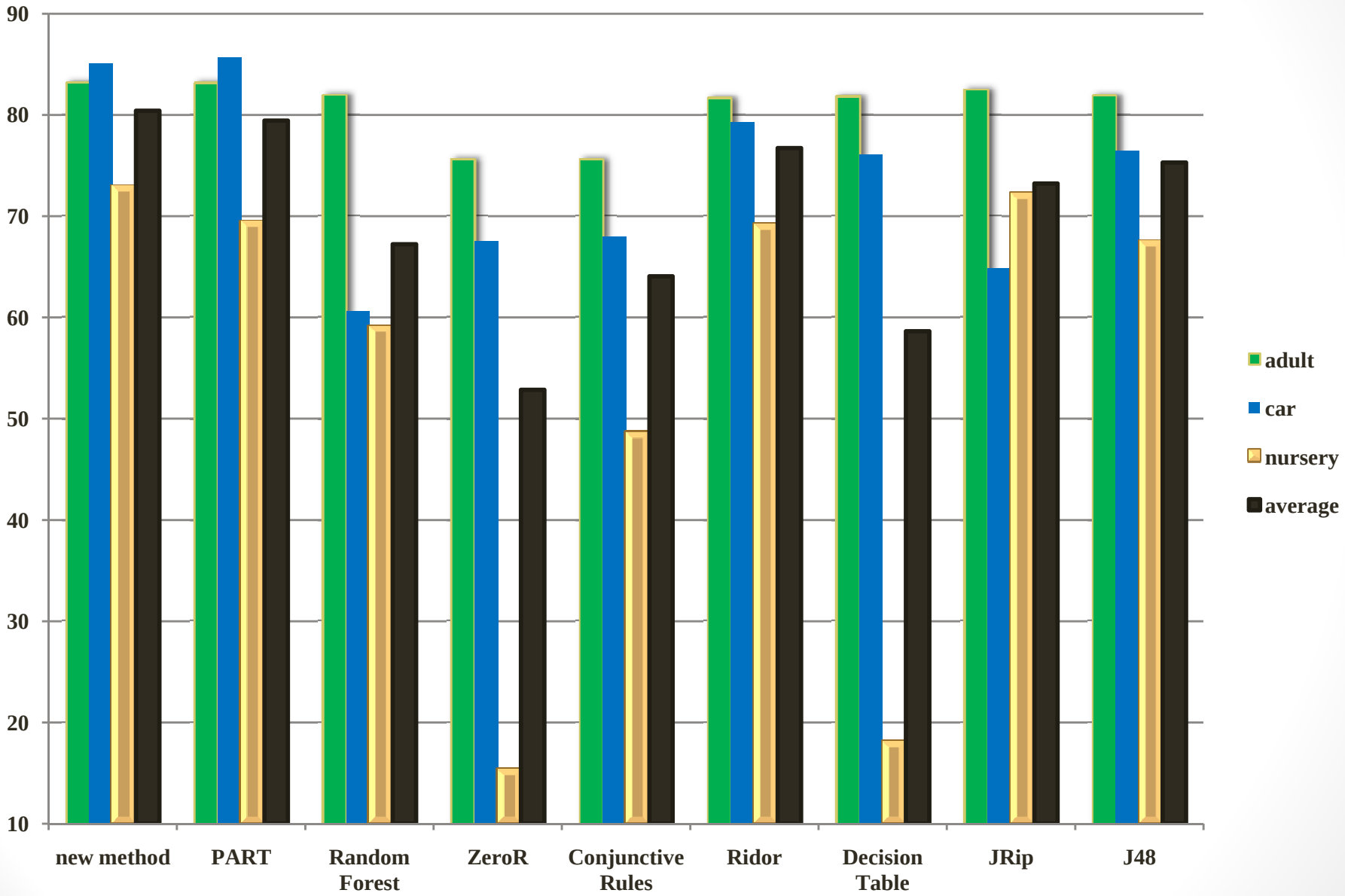
روش / تکرار	تکرار اول	تکرار دوم
صحت روش	۵۹.۲	۷۲.۹۵

- مقایسه‌ی نتایج

روش	روش پیشنهادی	روش PART	روش Random Forest	روش ZeroR	روش CR	روش Ridor	روش Decision table	روش JRip	روش J48
درستی	۷۲.۹۵	۶۹.۵	۵۹.۱	۱۵.۳۵	۴۸.۶۵	۶۹.۲	۱۸.۱	۷۲.۳	۶۷.۵۵

ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

مقایسه‌ی نتایج



ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

تحلیل و بررسی نتایج

- کارایی مطلوب در مورد دو مجموعه داده‌ی Adult و Nursery
  - ✓ استفاده از نتایج همه‌ی عامل‌ها و سازماندهی آنها در قالب یک مجموعه‌ی نهایی
  - ✓ بکارگیری یک روش مناسب برای رتبه‌بندی قوانین(با استفاده از تحلیل حوزه)
  - ✓ حذف قوانین زائد در پایان هر تکرار
- کارایی نامطلوب در مورد مجموعه داده‌ی Car
  - ✓ وابستگی روش مربوطه به مجموعه داده مورد بررسی
    - کم بودن حجم داده به منظور تحلیل حوزه و آزمایش نتایج
    - نامتوازن بودن مجموعه داده‌ی مورد بررسی

نتیجه‌گیری

- برتری‌ها
  - ✓ نبود مشکلات انتقال و ذخیره‌سازی داده
  - ✓ بکارگیری روش مربوطه بدون ایجاد اختلال در عملکرد سیستم موجود
  - ✓ انعطاف‌پذیری بیشتر در رتبه بندی و انتخاب قوانین(تحلیل حوزه با استفاده از یک مجموعه داده در همان حوزه)
  - ✓ در اختیار داشتن یک مجموعه‌ی بهینه از قوانین(رتبه‌بندی مناسب و انتخاب قوانین مفید)
- کاستی‌ها
  - ✓ وابستگی به مجموعه داده‌ی منبع

ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی

کارهای آینده

- استفاده از فاکتورهای بیشتر به منظور ارزیابی قوانین
- ارزیابی و انتخاب قوانین در عامل‌های سطح پایین
- پوشش انواع متفاوت از داده‌ها
  - ✓ بررسی داده‌ها به منظور کشف دنباله‌ای از وقایع(کاربرد در حوزه‌ی پزشکی)

مقالات مستخرج از پایان‌نامه

- دوستدار، م.، ناظمی، ا. و علیزاده، ح.، "ارائه‌ی یک روش جدید برای رده‌بندی مجموعه داده‌های ناهمگون با استفاده از مدل چند عاملی" ، کنفرانس فناوری اطلاعات و دانش (ارسال شده).

منابع منتخب

- Han, J. and Kamber,M., Data mining: concepts and techniques, Elsevier Pub., 2006.
- Cao,L., Data mining and multi-agent integration, Springer Science+Bussines Media, 2009.
- Gorodetsky,V., Autonomus intelligent systems: agents and data mining, Springer-Verlag, 2007.
- Rao,K. R.,Rao, M. N.,Prasad Y. S. andRamya, K. R., "Interaction and Integration of Agent Mining in Distributed Data Environment", In *International Journal of Computer Science & Emerging Technologies(E-ISSN: 2044-6004)*, Volume 1, Issue 4, 2010.
- Gao,J.,Denzinger, J. andJames, R. C., "A Cooperative Multiagent Data Mining Model and Its Application to Medical Data on Diabetes", In *Proc. Autonomous Intelligent Systems: Agents and Data Mining*, LNAI 3505, pp. 93–107, Springer-Verlag, 2005.
- Veloso, A. and Meira Jr, W.,"Rule generation and rule selection techniques for cost-sensitive associative classification." *Proc. of the Brazilian Symposium on Databases (SBBD)*. 2005.
- Wang, Y. J., Xin, Q. and Coenen, F.,"Hybrid rule ordering in classification association rule mining." *Transactions on Machine Learning and Data Mining* 1.1, pp. 1-16, 2008.
- Coenen, F. andLeng, P.,"An evaluation of approaches to classification rule selection." *Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on*. IEEE, 2004.
- Baralis, E., Chiusano, S. and Garza, P.,"On support thresholds in associative classification", In *Proceedings of the 2004 ACM symposium on Applied computing*, pp. 553-558. ACM, 2004.
- Todorovski, L. and Džeroski S., "Combining classifiers with meta decision trees." *Machine Learning* 50, pp. 223-249, 2003.

با سپاس فراوان از همراهی و توجه شما