# بسم الله الرحمن الرحيم

**دانشکده فنی و مهندسی**

**بخش مهندسی کامپیوتر**

**پایان‌نامه تحصیلی برای دریافت درجه کارشناسی ارشد**
**رشته مهندسی کامپیوتر گرایش هوش مصنوعی**

---

# کاربرد محاسبات دی.ان.ای در حوزه داده‌کاوی

---

**مؤلف:**

**رامین معاذالهی**

**استاد راهنما:**

**دکتر علی‌اکبر نیک‌نفس**

**استاد مشاور:**

**دکتر مجید محمدی**

**بهمن‌ماه ۱۳۹۱**

این پایان‌نامه به عنوان یکی از شرایط درجه کارشناسی ارشد به

**بخش مهندسی کامپیوتر**

**دانشکده فنی و مهندسی**

**دانشگاه شهید باهنر کرمان**

تسلیم شده است و هیچگونه مدرکی به عنوان فراغت از تحصیل دوره مزبور شناخته نمی‌شود.

دانشجو : رامین معاذالهی

استاد راهنما : دکتر علی‌اکبر نیک‌نفس

استاد مشاور : دکتر مجید محمدی

داور ۱ : دکتر فرامرز صادقی

داور۲ : دکتر حمید میروزیری

معاونت پژوهشی و تحصیلات تکمیلی دانشکده : دکتر مریم احتشام زاده

**تقدیم به** خدایی که آفرید جهان را، انسان را، عقل را، علم را، معرفت را، عشق را و به کسانی که عشقشان را در وجودم دمید.

**با سپاس از** سه وجود مقدس:

آنان که ناتوان شدند تا ما به توانایی برسیم...

موهایشان سپید شد تا ما روسفید شویم...

و عاشقانه سوختند تا گرمابخش وجود ما و روشنگر راهمان باشند...

پدرانمان

مادرانمان

استادانمان

## چکیده

در سال‌های اخیر قدرت محاسبات دی.ان.ای در حل مسائل پیچیده کامپیوتری به اثبات رسیده است. به دلیل قدرت پردازش موازی بسیار زیاد مولکول‌های دی.ان.ای، امکان جستجوی تمام فضای راه‌حل با استفاده از عملگرهای بیولوژیکی میسر می‌شود. در این تحقیق دو روش بر پایه محاسبات دی.ان.ای برای دسته‌بندی داده‌ها و استخراج قوانین همبستگی که در حوزه داده‌کاوی مطرح می‌شوند پیشنهاد می‌کنیم. در روش اول نشان می‌دهیم که چگونه می‌توان مدل دسته‌بندی داده‌ها را با استفاده از رشته-های دی.ان.ای تولید کرده و با اعمال توالی از عملگرهای بیولوژیکی روی مدل ایجادشده، دسته‌بندی داده‌های جدید را انجام داد. نتایج شبیه‌سازی روش پیشنهادی روی پنج مجموعه‌داده از UCI و مقایسه با روش C5.0، برتری روش پیشنهادی را از لحاظ دقت دسته‌بندی نشان می‌دهد. در روش دوم نشان می‌دهیم که چگونه می‌توان به طریق مشابه کلیه قوانین همبستگی را با استفاده از رشته‌های دی.ان.ای تولید کرده و با اعمال توالی از عملگرهای بیولوژیکی، قوانینی که حداقل پشتیبان و اطمینان تعیین‌شده را ارضا می‌کنند در زمان چندجمله‌ای استخراج کرد. لازم به ذکر است که استخراج کلیه قوانین همبستگی در کامپیوترهای سیلیکونی در بدترین حالت دارای پیچیدگی زمانی نمایی است، بنابراین مزیت روش پیشنهادی آشکار می‌شود.



## کلمات کلیدی:

محاسبات دی.ان.ای – داده‌کاوی – مدل آدلمان لیپتون – دسته‌بندی - استخراج قوانین همبستگی.

# فهرست مطالب

**عنوان** **صفحه**

# فهرست جداول

**شماره و عنوان** **صفحه**

# فهرست شکل ها

**شماره و عنوان** **صفحه**

# فصل اول:
# مقدمه

## ۱-۱ مقدمه

امروزه حجم عظیمی از داده‌های خام توسط شرکت‌ها و سازمآن‌ها در پایگاه داده‌ها ذخیره شده و روز به روز به میزان این داده‌ها افزوده می‌شود. اگر این داده‌ها به روش مناسبی پردازش شوند، اطلاعات مفیدی از آن‌ها استخراج می‌شود که در نهایت منجر به ارتقا سازمان و ارائه بهتر خدمات توسط آن‌ها می‌شود. روش‌های زیادی برای تحلیل داده‌ها و استخراج اطلاعات سودمند از آن‌ها وجود دارد که یکی از این روش‌ها داده‌کاوی[1] است.

بسیاری از تکنیک‌های مورد استفاده در داده‌کاوی نیازمند پردازش‌های زیاد توسط سیستم کامپیوتری است که در برخی موارد خارج از توان سیستم‌های موجود می‌باشد. از سوی دیگر در سال‌های اخیر، توانایی سیستم‌های بیولوژیکی در حل مسائل ان.پی‌کامل[2] به اثبات رسیده است. در این سیستم‌ها با بهره‌گیری از قابلیت پردازش موازی بسیار زیاد مولکول‌های دی.ان.ای[3]، کلیه محاسبات در لوله‌های آزمایشی انجام می‌پذیرد که منجر به ظهور علم میان‌رشته‌ای جدیدی بنام محاسبات دی.ان.ای[4] شده

[1] data mining
[2] NP-complete
[3] DNA (deoxyribonucleic acid)
[4] DNA computing

است. به دلیل قدرت پردازش موازی بسیار زیاد مولکول‌های دی.ان.ای می‌توان از آن‌ها برای حل مسائل پیچیده کامپیوتری استفاده کرد.

علی‌رغم قدرت زیاد محاسبات دی.ان.ای هنوز پیوند مناسبی میان این حوزه و حوزه داده‌کاوی ایجاد نشده است. از اینرو در این تحقیق می‌کوشیم تا کاربرد محاسبات دی.ان.ای برای دسته‌بندی داده‌ها و استخراج قوانین همبستگی از حوزه داده‌کاوی را بررسی کنیم. در ادامه فصل اول ضمن معرفی اجمالی داده‌کاوی و محاسبات دی.ان.ای به برخی از روش‌هایی که تاکنون در مقالات برای دسته‌بندی و استخراج قوانین همبستگی پیشنهاد شده است اشاره ای داریم. همچنین نگاهی اجمالی به برخی مقالات مهم در زمینه محاسبات دی.ان.ای داریم. در فصل دوم به داده‌کاوی و برخی تکنیک‌های آن می‌پردازیم. محاسبات دی.ان.ای به همراه کاربردها، مزایا، انواع مدل‌ها و مسائل نمونه به صورت کامل در فصل سوم شرح داده می‌شود. در فصل چهارم، دو روش، یکی برای حل مساله دسته‌بندی و دیگری برای استخراج قوانین همبستگی در مساله تحلیل سبد خرید، با استفاده از محاسبات دی.ان.ای پیشنهاد می‌کنیم. در نهایت فصل پنجم به ارزیابی روش‌های پیشنهادی و نتایج مهمی که از این تحقیق حاصل می‌شود اختصاص دارد. در فصل پنجم همچنین پیشنهاداتی برای ادامه تحقیقات در این زمینه ارائه می‌شود.

## ۱-۲ داده‌کاوی

داده‌کاوی یا کشف‌دانش فرآیندی است که طی آن داده‌ها از جنبه‌های گوناگون تجزیه و تحلیل شده و بصورت اطلاعات مفید-اطلاعاتی که می‌توانند باعث افزایش درآمد و یا کاهش هزینه شوند- خلاصه می‌شوند. داده‌ها می‌توانند واقعیات، اعداد و یا متن‌هایی باشند که بوسیله کامپیوتر قابل پردازش باشند. امروزه سازمآن‌ها حجم وسیعی از داده‌هایشان را در قالب شکل‌های متفاوت و در پایگاه داده‌های گوناگون ذخیره می‌کنند. به عنوان مثال داده‌ها می‌توانند لیست فروش، هزینه و یا سرمایه‌گذاری در یک سازمان باشند.

حاصل پردازش داده‌ها، اطلاعات می‌باشد که بیانگر الگوها، وابستگی‌ها و یا ارتباط بین داده‌هاست. به عنوان مثال، تحلیل داده‌های فروش یک فروشگاه می‌تواند اطلاعات سودمندی از قبیل نحوه چیدمان کالاها به صورتیکه احتمال فروش آن‌ها با هم بیشتر باشد را در اختیار فروشنده قرار دهد.

از جمله کارهایی که در حوزه داده‌کاوی قابل انجام است می‌توان به دسته‌بندی[1]، خوشه‌بندی[2]، استخراج قوانین همبستگی[3]، تخمین[4] و پیش‌بینی[5] اشاره کرد. در دسته‌بندی مجموعه‌ای از داده‌ها را در اختیار داریم که در دسته‌های متفاوت قرار داده شده‌اند. هدفِ دسته‌بندی، پیش‌بینی دسته صحیح برای هر داده است که این کار از طریق پیدا کردن قوانین دسته‌بندی[6] صورت می‌پذیرد. هر قانون بیانگر رابطه منطقی بین داده‌ها و دسته‌ای است که داده‌ها در آن قرار دارند. پس از استخراج قوانین دسته‌بندی، می‌توان از آن‌ها برای دسته‌بندی داده‌های جدید که دسته آن‌ها مشخص نیست استفاده کرد. خوشه‌بندی نیز مانند دسته‌بندی است با این تفاوت که فقط داده‌ها را در اختیار داریم و دسته‌بندی داده‌ها بایستی با توجه به میزان شباهتشان به یکدیگر صورت بپذیرد. در بحث استخراج قوانین همبستگی، هدف پیدا کردن روابط همبستگی و الگوهای تکرارشونده بین مجموعه‌ای از کالاها در پایگاه داده‌های تراکنشی است. با استفاده از قوانین استخراج‌شده، می‌توان سبدهای خرید کالا را بگونه‌ای طراحی کرد که بیشترین شانس فروش را داشته باشند. در تخمین و پیش‌بینی که شبیه دسته‌بندی می‌باشند، هدف تعیین یک مقدار عددی به عنوان خروجی برای هر داده می‌باشد. تفاوت پیش‌بینی با تخمین این است که در پیش‌بینی نتیجه مربوط به آینده می‌باشد، به عنوان مثال پیش‌بینی نرخ طلا در سه ماهه آینده سال جاری.

## ۱-۳ محاسبات دی.ان.ای

محاسبات دی.ان.ای شیوه‌ای بدیع از محاسبات است که در آن از مولکول‌های دی.ان.ای بجای مدارات منطقی دیجیتال استفاده می‌شود. در این شیوه محاسبات، هر سلول بیولوژیکی مانند یک کامپیوتر پیچیده با توان پردازشی بسیار بالاست. الفبای محاسبات در این روش، بازهای آمینواسیدی بوده که با حروف A، T، C و G نمایش داده شده و اجزای اصلی تشکیل‌دهنده مولکول‌های دی.ان.ای می‌باشند. برای کد کردن اطلاعات از حروف ذکر شده بجای صفر و یک رایج در کامپیوترهای الکترونیکی استفاده می‌شود. مزیت اصلی محاسبات دی.ان.ای توان پردازش موازی بسیار زیاد مولکول‌های دی.ان.ای است که امکان حل مسائل پیچیده و بخصوص مسائل ان.پی‌کامل را که در برخی موارد از توان پردازشی کامپیوترهای الکترونیکی خارج است، فراهم می‌کند.

---

[1] classification
[2] clustering
[3] association rule mining
[4] estimation
[5] prediction
[6] classification rules

مسائل ان.پی‌کامل دسته‌ای از مسائل تصمیم‌گیری هستند که هیچ راه‌حل سریعی برای آن‌ها وجود ندارد. به عبارت دیگر با افزایش اندازه ورودی این مسائل، زمان مورد نیاز برای حل آن‌ها با استفاده از الگوریتم‌های فعلی بصورت نمایی رشد می‌کند. این مسائل نمونه‌های مناسبی برای حل با استفاده از محاسبات دی.ان.ای می‌باشند، زیرا با استفاده از قدرت پردازش موازی مولکول‌های دی.ان.ای امکان بررسی تمام جواب‌های ممکن و غیرممکن برای مساله در زمان قابل قبول امکان‌پذیر است. مساله مسیر همیلتونی[۱] [۱]، فروشنده دوره‌گرد[۲] [۲,۳]، رنگ‌آمیزی گراف[۳] [۴,۵]، مسئله گروهگ[۴] [۶]، مجموعه مستقل[۵] [۷]، افزار مجموعه[۶] [۸]، مساله کوله‌پشتی[۷] [۹]، مساله ارضا قیدهای منطقی[۸] [۱۰] و مساله همریختی گراف[۹] [۱۱] از جمله مسائل ان.پی می‌باشند که راه‌حل آن‌ها با استفاده از محاسبات دی.ان.ای ارائه شده است.

## ۱-۴ مروری بر کارهای پیشین

در بحث دسته‌بندی داده‌ها هدف ساخت مدل مناسبی است که بتوان از آن برای دسته‌بندی داده‌های جدید استفاده کرد. کوینلان [۱۲] در سال ۱۹۸۶ الگوریتم ID3 را برای ساخت مدل دسته‌بندی بر اساس درخت‌های تصمیم ارائه داد. در درخت ساخته‌شده هر مسیر از ریشه به برگ معرف یک قانون دسته‌بندی است. برای ساخت درخت تصمیم، از جستجوی حریصانه بالا به پایین در فضای درخت‌های موجود استفاده می‌شود. بدین صورت که در هر مرحله آن ویژگی که به بهترین شکل داده‌ها را از هم جدا می‌کند به عنوان گره بعدی درخت انتخاب می‌شود. معیار انتخاب بهترین ویژگی نیز بر اساس مقداری آماری به نام بهره‌اطلاعات صورت می‌پذیرد. الگوریتم‌های C4.5 و C5.0 نیز نسخه‌های تکامل‌یافته الگوریتم ID3 می‌باشند. از مزیت‌های درخت‌های تصمیم می‌توان به عدم حساسیت به داده‌های نویزی و کارآمد بودن آن برای حجم زیاد داده‌ها اشاره کرد اما از آنجاییکه ساخت درخت به روش حریصانه انجام می‌شود، بنابراین تضمینی برای یافتن درخت بهینه وجود ندارد.

در سال ۱۹۹۶ لو و همکاران [۱۳] روشی برای استخراج قوانین دسته‌بندی با استفاده از شبکه عصبی مصنوعی[۱۰] پیشنهاد دادند. در روش پیشنهادی آن‌ها، ابتدا یک شبکه عصبی سه لایه ساخته‌شده و با

---

[1] hamiltonian path problem
[2] traveling salesman problem
[3] vertex coloring
[4] clique problem
[5] independent set problem
[6] set splitting problem
[7] knapsack problem
[8] satisfiability problem
[9] graph isomorphism
[10] artificial neural network

استفاده از داده‌های آموزشی، آموزش داده می‌شود تا به دقت مورد نظر برسد. سپس با استفاده از یک الگوریتم هرس‌بندی، اتصالات و گره‌های اضافی به گونه‌ای هرس می‌شوند که دقت دسته‌بندی کاهش نیابد که منجر به تولید شبکه عصبی ساده تری می‌شود. در مرحله بعد قوانین دسته‌بندی از این شبکه استخراج می‌شوند. یکی از معایب این روش زمان نسبتاً زیادی است که برای آموزش شبکه عصبی صرف می‌شود. نویسندگان مقاله برای رفع این مشکل دو روش پیشنهاد می‌دهند که یکی استفاده از آموزش تدریجی و دیگری کاهش تعداد ورودی‌های شبکه عصبی با استفاده از الگوریتم‌های انتخاب ویژگی[1] است.

در سال ۲۰۰۲ پارپینلی و همکاران [۱۴] روشی بنام اَنت‌ماینر برای استخراج قوانین دسته‌بندی با استفاده از الگوریتم اجتماع مورچه ارائه دادند. در روش پیشنهادی آن‌ها از معیار بی‌نظمی[2] به عنوان معیار شهودتی[3] در تصمیم‌گیری مورچه‌ها برای انتخاب میان صفات استفاده می‌شود. هر مسیری که توسط یک مورچه طی شود منجر به تولید یک قانون دسته‌بندی می‌شود. پس از ساخت قانون توسط مورچه، آن قانون هرس شده که باعث افزایش دقت، سادگی و قابل فهم‌تر شدن قانون می‌شود. لیو و همکاران [۱۵] با ارائه معیار شهودتی ساده‌تری که بر پایه تخمین چگالی کار می‌کرد، نسخه دوم اَنت‌ماینر را ارائه کردند. ایده اصلی آن‌ها در استفاده از معیار ساده‌تر این بود که لازم نیست معیار شهودتی خیلی دقیق باشد. نسخه دوم اَنت ماینر نسبت به اولین نسخه پیچیدگی محاسباتی کم‌تر و دقت دسته‌بندی تقریباً یکسانی را ارائه می‌داد. در سال ۲۰۰۴ سومین نسخه اَنت‌ماینر توسط لیو و همکاران [۱۶] ارائه شد که در آن استراتژی بروزرسانی فرمون‌ها و قانون انتقالِ حالت بهبود یافته بودند که باعث افزایش دقت دسته‌بندی نسبت به نسخه‌های پیشین شد.

در سال ۲۰۰۳ روشی برای استخراج قوانین دسته‌بندی با استفاده از برنامه‌نویسی عبارات ژنی[4] توسط ژو و همکاران [۱۷] ارائه شد. برنامه‌نویسی عبارات ژنی که ترکیب الگوریتم ژنتیک[5] و برنامه‌نویسی ژنتیک[6] است، ارتباط بین صفات را به صورت روابط ریاضی استخراج می‌کند. در روش پیشنهادی آن‌ها برای حل مساله دسته‌بندی با $n$ دسته از تکنیک یادگیری یکی در مقابل همه[7] استفاده می‌شود که مساله را به $n$ مساله دسته‌بندی باینری تبدیل کرده که با استفاده از برنامه‌نویسی عبارات ژنی قابل حل

---

[1] feature selection
[2] entropy
[3] heuristic
[4] gene expression programming
[5] genetic algorithm
[6] genetic programming
[7] one-against-all learning

می‌باشند. همچنین تابع ارزیابی بگونه‌ای تعریف می‌شود که برای هر قانون، هم کامل‌بودن و هم سازگاربودن در نظر گرفته می‌شوند. از مزایای روش ارائه شده عدم حساسیت آن به داده‌های نویزی به دلیل استفاده از اصل حداقل طول توصیفی[1] و هرس کردن مجموعه قوانین استخراج‌شده است. هرس کردن قوانین در دو مرحله انجام می‌شود: یکی در حین یادگیری قوانین که از ایجاد قوانین پیچیده (که معمولاً بدلیل داده‌های نویزی ایجاد می‌شوند) جلوگیری کرده و دیگری پس از فرآیند یادگیری است که برای حذف تضاد بین قوانین لازم است. مشکل عمده روش‌هایی که از برنامه‌نویسی عبارات ژنی استفاده می‌کنند این است که در برخی موارد قوانین تولیدشده بسیار پیچیده بوده و فهم آن‌ها برای ما دشوار است.

چیو [۱۸] در سال ۲۰۰۵ از الگوریتم ژنتیک مبتنی بر قیود برای استخراج قوانین دسته‌بندی استفاده کرد. از آنجا که هر داده مجموعه‌ای از صفات است، در روش پیشنهادی می‌توان قیود را بصورت رابطه بین این صفات بیان کرد. همچنین برای تولید کروموزوم هایی که همه قیود را ارضا کنند از استدلال مبتنی بر قیود استفاده شده است. از آنجا که با افزایش تعداد صفات، تعداد قوانین دسته‌بندی ممکن بصورت نمایی افزایش می‌یابد، استفاده از الگوریتم ژنتیک برای جستجوی جهت‌یافته این فضای بزرگ مناسب به نظر می‌رسد.

از دیگر مطالعات انجام‌شده در زمینه داده‌کاوی می‌توان به استخراج قوانین همبستگی اشاره کرد. یکی از کاربردهای مهم این قوانین استفاده از آن‌ها در سیستم‌های توصیه‌گر است. فرآیند استخراج قوانین همبستگی از دو مرحله تشکیل می‌شود: [۱۹] در مرحله اول سبد کالاهای مکرر–سبد کالاهایی که تعداد تکرارشان در کل پایگاه داده از مقدار مشخصی بیشتر است– شناسایی شده و سپس در مرحله دوم قوانین همبستگی با استفاده از این سبد کالاها تولید می‌شوند. از آنجاکه مرحله دوم ساده و سرراست است، بیشتر تحقیقات انجام‌شده در زمینه استخراج قوانین همبستگی در رابطه با مرحله اول یعنی شناسایی سبد کالاهای مکرر بوده است.

آگراوال و همکاران [۱۹] در سال ۱۹۹۳ اولین الگوریتم را برای استخراج قوانین همبستگی بنام الگوریتم اِیز[2] ارائه دادند. در این الگوریتم برای شناسایی سبد کالاهای مکرر[3]، کل پایگاه داده

[1] minimum description length
[2] AIS (Agrawal, Imielinski, Swami)
[3] frequent itemset

چندین‌بار پیمایش می‌شود که یکی از معایب این الگوریتم هم به شمار می‌رود. در اولین پیمایش پایگاه داده، کل سبد کالاهای مکرر تک‌عضوی شناسایی شده و با اضافه کردن کالاهای باقیمانده به هر کدام از این سبدها، سبد کالاهای دو عضوی منتخب ساخته می‌شوند که با پیمایش مجدد پایگاه داده، سبد کالاهای مکرر دوعضوی از میان کل سبد کالاهای منتخب شناسایی می‌شوند. مشکل اصلی الگوریتم اِیز این است که بسیاری از سبد کالاهایی که امکان تبدیل‌شدن به سبد کالاهای مکرر را ندارند بیهوده تولید شده و زمان و فضای زیادی برای پردازش آن‌ها تلف می‌شود؛ هر چند در این الگوریتم روش‌هایی برای هرس‌کردن سبد کالاهای منتخب و همچنین مدیریت حافظه پیشنهاد شده است.

در سال ۱۹۹۴ آگراوال و سریکانت [۲۰] الگوریتم اپیریوری[1] را ارائه دادند تا مشکلات موجود در الگوریتم اِیز را برطرف سازند. پیمایش بسیار زیاد پایگاه داده و تولید سبد کالاهای منتخب غیرضروری از مشکلات الگوریتم اِیز می‌باشند که در الگوریتم اپیریوری به مقدار قابل توجهی کاهش یافته‌اند. در این الگوریتم تکنیک تولید سبد کالاهای منتخب و هرس‌کردن بهبود داده شده‌اند که همین امر باعث افزایش کارایی آن در مقایسه با الگوریتم اِیز شده است. روش تولید سبد کالاهای منتخب بدین صورت است که ابتدا با یکبار پیمایش پایگاه داده، مجموعه تمام سبد کالاهای مکرر تک‌عضوی شناسایی شده و از طریق پیوندزدن مجموعه حاصل با خودش، سبد کالاهای منتخب دوعضوی ایجاد می‌شوند. در هر مرحله مجموعه‌ی سبد کالاهای منتخب $k$ عضوی با خودش پیوند زده شده و مجموعه‌ی سبد کالاهای منتخب $k+1$ عضوی ساخته می‌شود. همچنین با استفاده از این ویژگی که تمام زیرمجموعه‌های $k$ عضوی از یک سبد کالای مکرر $k+1$ عضوی، خودشان سبد کالای مکرر می‌باشند، تعداد سبد کالاهای منتخب k عضوی کاهش می‌یابد. بدین ترتیب تعداد پیمایش‌های پایگاه داده به نسبت زیادی کاهش می‌یابد. با این وجود هنوز هم پیمایش کل پایگاه داده به تعداد زیاد و تکنیک پیچیده تولید مجموعه سبد کالاهای منتخب از مشکلات اصلی الگوریتم اپیریوری است. انتخاب بخشی از پایگاه داده بجای کل آن برای پیمایش، نمونه‌برداری از داده‌ها، استفاده از تکنیک‌های هرس‌کردن متفاوت مجموعه سبد کالاهای منتخب از جمله اصلاحاتی است که در مطالعات مختلف برای افزایش کارایی الگوریتم اپیریوری پیشنهاد شده است. از جمله این الگوریتم ها می‌توان به [۲۰,۲۱,۲۲] اشاره کرد.

[1] apriori