دانشکده فنی و مهندسی

گروه صنایع

پایان‌نامه کارشناسی ارشد رشته‌ی مهندسی صنایع گرایش مهندسی صنایع

# ارائه‌ی یک مدل خوشه‌بندی تودرتو با رویکرد مدل‌های برنامه‌ریزی ریاضی

**استاد راهنما:**

دکتر محمد مهدی سپهری

**استاد مشاور:**

دکتر عباس حاج فتحعلیها

**نگارش:**

نفیسه سلیمانی

پاییز ۱۳۸۷

**کلیه‌ی حقوق این پایان‌نامه**

**متعلق به دانشگاه شاهد است.**

تقدیم به

آن خورشید آشکار نهان

و تقدیم به

پدرم که درس استقامت و تلاش را به من آموخت

و مادرم به خاطر تلاش، محبت، مهر و خوبی اش

و عطیه جان،

و همه آنانی که از موفقیت من دلشاد می شوند.

# تشکر و قدردانی

پروردگار یکتا را سپاسگزارم که به من توان به پایان رساندن مرحله‌ای دیگر از زندگیم را عطا فرمود. اکنون که در پایان این راه و آغاز مرحله‌ای دیگر قرار دارم بر خود لازم می‌دانم که از زحمات کلیه اساتید بزرگواری که مرا در رسیدن به این مرحله یاری داده‌اند سپاسگزاری نمایم.

بدینوسیله مراتب قدردانی خود را از جناب آقای دکتر محمد مهدی سپهری که به عنوان استاد راهنما، بار، رهنمودها و دلسوزی‌های خود سهم عمده‌ای در انجام این پایان‌نامه داشته‌اند، اعلام می‌دارم و از جناب آقای دکتر عباس حاج فتحعلیها که مشاوره‌ی اینجانب را در این پروژه بر عهده داشته‌اند نیز سپاسگزارم.

از جناب آقای دکتر حمید فرورش که با اهتمام فراوان من را در آماده‌سازی این پایان نامه و تکمیل این پژوهش یاری نموده و بدینوسیله بر غنای آن افزودند کمال تشکر و قدردانی را دارم.

از پدر و مادر مهربانم که راه دانش آموزی را پیش روی من نهادند و اولین معلم من بودند بسیار سپاسگزارم.

## چکیده

خوشه‌بندی، قرار دادن داده‌ها در کنار یکدیگر، بر اساس ماهیت درونی آن‌هاست. محققان طبقه‌بندی‌های مختلفی برای الگوریتم‌های خوشه‌بندی ارائه داده‌اند. یک طبقه‌بندی سه‌تایی شامل مدل‌های خوشه‌بندی افرازی مدل‌های خوشه‌بندی سلسله‌مراتبی و خوشه‌بندی فازی، از کلی‌ترین تقسیم‌بندی‌ها محسوب می‌شود. به طوری‌که مدل‌های افرازی، سلسله‌مراتبی و فازی اشتراکی با یکدیگر ندارند. با این حال، استفاده از الگوریتم‌های سلسله‌مراتبی به عنوان جواب اولیه برای خوشه‌بندی افرازی، اولین پل ارتباطی بین حوزه‌ی خوشه‌بندی افرازی و خوشه‌بندی سلسله‌مراتبی محسوب می‌شود.

در این پژوهش یک مدل خوشه‌بندی تودرتو، از ترکیب حوزه‌ی خوشه‌بندی افرازی و حوزه‌ی سلسله‌مراتبی ارائه می‌شود. این مدل توسط ابزار بهینه‌سازی، می‌تواند داده‌ها را در چند سطح متداخل، پیوسته و همزمان، خوشه‌بندی نماید. کاربرد مدل خوشه‌بندی تودرتو، در تعیین خوشه‌های شبیه به هم و تشخیص میزان نزدیکی آن‌ها به یکدیگر است و می‌توان آن را دومین پل ارتباطی بین حوزه‌ی خوشه‌بندی افرازی و سلسله‌مراتبی دانست. مدل خوشه‌بندی تودرتو، بر مبنای خوشه‌بندی افرازی ساخته شده است. به همین دلیل سعی بر آن شد از تابع هدفی استفاده شود که داده‌ها را با درصد خلوص بالاتری خوشه‌بندی نماید. تابع هدف مورد استفاده در این ساختار، با رویکرد خارج از خوشه و حداکثر کردن مجموع فواصل بین خوشه‌ها شکل گرفت که این تابع هدف برای اولین بار در این پژوهش ارائه می‌شود. مدل خوشه‌بندی تودرتوی ارائه شده را K-NCLUS نامیده‌ایم و دارای دو ویژگی اساسی است. اول آن‌که سطوح پایین خوشه‌بندی در سطوح بالاتر شکسته نمی-شود و دوم آن‌که کلیه‌ی سطوح خوشه‌بندی به صورت همزمان خوشه‌بندی می‌شود. مدل K-NCLUS بر اساس اصل استقرا، راه‌حلی برای حالت‌های اولیه‌ی یک‌سطحی، دوسطحی، سه‌سطحی ارائه و سپس بر پایه‌ی دو اصل اساسی تا سطح $k$ام تعمیم داده می‌شود و یک ساختار کلی برای مسئله ایجاد می‌نماید. برای اثبات درستی مدل ارائه شده نیز چند پایگاه داده‌ی مجازی مورد استفاده قرار می‌گیرد. کلیه‌ی سطوح مدل خوشه‌بندی تودرتو دارای عملکرد یکسانی است که با ماهیت سلسله‌مراتبی به یکدیگر مربوط شده‌اند. مدل تک‌سطحی خوشه‌بندی تودرتو، یک مدل خوشه‌بندی افرازی است که با رویکرد جداسازی خوشه‌ها از یکدیگر و با تابع هدف حداکثر کردن مجموع فواصل بین خوشه‌ها ساخته می‌شود و سطوح بالاتر خوشه‌بندی تودرتو، روی آن مدل تک‌سطحی قرار می‌گیرد. علاوه بر آن، یک مدل خوشه‌بندی افرازی دیگری با رویکرد فشردگی درون خوشه و تابع هدف حداقل کردن مجموع فواصل درون خوشه ارائه می‌شود. به ترتیب، مدل اول را مدل BGSD و مدل دوم را مدل WGSD نامیده‌ایم. به منظور اعتبارسنجی درونی و بیرونی مدل‌های ارائه شده از سه پایگاه داده‌ی واقعی و سه مجموعه داده‌ی مجازی استفاده شده است. برای مقایسه‌ی مدل‌های پیشنهادی با مدل‌های ارائه شده در منابع تحقیق نیز معیار مجموع مربعات درون خوشه در نظر گرفته شده است. نتایج بدست آمده نشان می‌دهد مدل BGSD داده‌ها را با خلوص بیشتر و مجموع مربعات فاصله‌ی کمتری خوشه‌بندی می‌کند.



**کلید واژگان**: خوشه‌بندی افرازی، خوشه‌بندی تودرتو ، برنامه‌ریزی ریاضی، خوشه‌بندی از پایین به بالا

## فهرست مطالب

**عنوان** **صفحه**


**چکیده** **ت**

**۱ فصل اول: ورودی به پژوهش و شناخت آن** **۱**

۱-۱ سرآغاز فصل نخست ..... ۲

۱-۲ تعریف مسئله و بیان موضوع پژوهش ..... ۲

۱-۳ سؤالات پژوهش ..... ۳

۱-۴ اهداف پژوهش ..... ۳

۱-۵ ضرورت و لزوم پژوهش ..... ۴

۱-۶ کاربردهای مدل ..... ۴

۱-۷ مرور کلی بر روش‌شناسی پژوهش ..... ۴

۱-۸ تعاریف و اصلاحات پایه ..... ۶

۱-۹ ساختار کلی پایان‌نامه ..... ۷

۱-۱۰ سرانجام فصل نخست ..... ۷

**۲ فصل دوم: بررسی پیشینه‌ی پژوهش** **۸**

۲-۱ سرآغاز فصل دوم ..... ۹

۲-۲ ادبیات موضوع ..... ۹

۲-۲-۱ طبقه‌بندی الگوریتم‌های خوشه‌بندی ..... ۱۲

۲-۲-۲ توابع هدف به کار رفته در مدل‌های خوشه‌بندی ..... ۲۰

۲-۲-۳ سایر الگوریتم‌های خوشه‌بندی ..... ۲۵

۲-۳ ارتباط بین تحقیقات گذشته و موضوع پژوهش ..... ۳۱

۲-۴ سرانجام فصل دوم ..... ۳۱

**۳ فصل سوم: مدل BGSD و مدل WGSD و محاسبات و واکافت نتایج روی داده‌های واقعی و مجازی** **۳۲**

۳-۱ سرآغاز فصل سوم ..... ۳۳

۳-۲ مدل غیرخطی BGSD ..... ۳۴

۳-۳ مدل خطی BGSD ..... ۳۵

۳-۴ مدل غیرخطی WGSD ..... ۳۶

۳-۵ مدل خطی WGSD ..... ۳۷

۳-۶ مقایسه‌ی مدل BGSD و مدل WGSD با مدل‌های k-means و MWGD ..... ۳۷

۳-۶-۱ نتایج حاصل از داده‌های مجازی بر روی مدل‌های BGSD، WGSD، k-means و MWGD ..... ۳۸

۳-۶-۱-۱ پایگاه داده‌ی مجازی ۸۰تایی، خوشه‌بندی، بررسی و تحلیل نتیجه ..... ۳۸

۳-۶-۱-۲ پایگاه داده‌ی مجازی ۹۰تایی، خوشه‌بندی، بررسی و تحلیل نتیجه ..... ۴۱

۳-۶-۱-۳ پایگاه داده‌ی مجازی ۲۱۶تایی، خوشه‌بندی، بررسی و تحلیل نتیجه ..... ۴۲

۳-۶-۱-۴ بررسی نموداری مجموع مربعات درون خوشه روی داده‌های مجازی ..... ۴۴

۳-۶-۲ نتایج حاصل از داده‌های واقعی بر روی مدل‌های BGSD،WGSD، k-means و MWGD ..... ۴۵
۳-۶-۲-۱ پایگاه داده‌ی Iris، خوشه‌بندی، اعتبارسنجی درونی/ برونی و تحلیل نتیجه ..... ۴۶
۳-۶-۲-۲ پایگاه داده‌یOriginal Owner and Donor، خوشه‌بندی، اعتبارسنجی درونی/برونی و تحلیل نتیجه..... ۴۶
۳-۶-۲-۳ پایگاه داده‌ی Glass، خوشه‌بندی، اعتبارسنجی درونی/ برونی و تحلیل نتیجه ..... ۴۷
۳-۶-۳ تحلیل نموداری درصد خلوص مدل‌ها بر روی داده‌های واقعی ..... ۴۸
۳-۷ آنالیز حساسیت مدل BGSD و تحلیل نموداری مدل خطی و غیرخطی آن ..... ۵۰
۳-۸ سرانجام فصل سوم ..... ۵۴
**۴ فصل چهارم: مدل خوشه‌بندی تودرتو K-NCLUS و محاسبات و واکافت نتایج روی داده‌های مجازی ۵۶**
۴-۱ سرآغاز فصل چهارم..... ۵۷
۴-۲ مدل K-NCLUS تغییر یافته‌ی مدل BGSD..... ۵۷
۴-۳ مدل غیرخطی 1-NCLUS ..... ۵۸
۴-۴ مدل خطی 1-NCLUS ..... ۵۹
۴-۵ مدل 2-NCLUS..... ۶۰
۴-۵-۱ مدل غیرخطی 2-NCLUS ..... ۶۰
۴-۵-۲ مدل گسترده‌ی 2-NCLUS ..... ۶۱
۴-۵-۳ تفاوت مدل 2-NCLUS و مدل گسترده‌ی 2-NCLUS..... ۶۲
۴-۵-۴ مدل خطی گسترده‌ی 2-NCLUS..... ۶۳
۴-۵-۵ نتایج حاصل از خوشه‌بندی دوسطحی بر روی داده‌های مجازی توسط مدل 2-NCLUS..... ۶۴
۴-۶ مدل 3-NCLUS..... ۶۷
۴-۶-۱ مدل غیرخطی 3-NCLUS ..... ۶۷
۴-۷ نتیجه‌ی حاصل از خوشه‌بندی سه‌سطحی روی داده‌های مجازی توسط مدل3-NCLUS ..... ۶۹
۴-۸ تعمیم الگوریتم Nested hierarchical clustering برای ساخت مدل K-NCLUS..... ۷۱
۴-۹ مقایسه‌ی بین مدل K-NCLUS و روش Bi-Section Clustering..... ۷۲
۴-۱۰ توسعه‌ی مدل K-NCLUS با رویکرد تعویض تابع هدف ..... ۷۳
۴-۱۰-۱مدل سه‌سطحی دو تابع هدفی ..... ۷۴
۴-۱۰-۲مدل سه‌سطحی یک تابع هدفی..... ۷۶
۴-۱۰-۳مدل سه‌سطحی سه تابع هدفی ..... ۷۸
۴-۱۱ سرانجام فصل چهارم ..... ۷۹
**۵ فصل پنجم: سرانجام پژوهش و گشایش افق‌های نو ۸۱**
۵-۱ سرآغاز فصل پنجم..... ۸۲
۵-۲ مروری بر فصول گذشته ..... ۸۲
۵-۳ دستاوردهای پژوهش ..... ۸۳
۵-۴ نوآوری‌های پژوهش ..... ۸۳

۵-۵ چگونگی و ساز و کارهای به کارگیری و پیاده‌سازی نتایج ..... ۸۴
۵-۶ تبیین و توجیه افق‌های پژوهش و مباحث مطالعاتی پیشنهادی ..... ۸۵
۵-۷ سرانجام فصل پایانی ..... ۸۵
**۶ پیوست الف (واژه‌نامه فارسی به انگلیسی) ۸۶**
**۷ پیوست ب (واژه‌نامه انگلیسی به فارسی) ۸۸**
**۸ منابع و مؤاخذ ۹۰**

## فهرست شکل‌ها

**عنوان** **صفحه**


شکل ۲-۱. مراحل کشف دانش / داده‌کاوی (Bradley et al, 1999) ..... ۲۶
شکل ۲-۲. چهار مرحله‌ی خوشه‌بندی (Rui, 2005) ..... ۲۹
شکل ۳-۱. مجموعه داده‌های اولیه ۸۰ تایی ..... ۳۹
شکل ۳-۲. مجموعه داده‌های تفکیک شده‌ی ۸۰ تایی ..... ۳۹
شکل ۳-۳. مجموعه داده‌های تفکیک شده به همراه داده‌های تصادفی ..... ۳۹
شکل ۳-۴. خوشه‌بندی داده‌های ۸۰ تایی توسط روش BGSD ..... ۴۰
شکل ۳-۵. پایگاه داده‌ی ۹۰تایی ..... ۴۱
شکل ۳-۶. خوشه‌بندی داده‌های ۹۰ تایی به روش BGSD ..... ۴۱
شکل ۳-۷. مجموعه داده‌ی ۲۱۶تایی ..... ۴۳
شکل ۳-۸. خوشه‌بندی داده‌های ۲۱۶تایی به روش BGSD ..... ۴۳
شکل ۴-۱. مجموعه داده‌های اولیه ۳۲تایی ..... ۶۵
شکل ۴-۲. خوشه‌بندی داده‌های اولیه در دو سطح متداخل توسط مدل 2-NCLUS ..... ۶۵
شکل ۴-۳. مجموعه داده‌های اولیه ۸۴تایی ..... ۶۶
شکل ۴-۴. خوشه‌بندی داده‌های اولیه در دو سطح متداخل توسط مدل 2-NCLUS ..... ۶۶
شکل ۴-۵. مجموعه داده‌های اولیه ۸۴تایی ..... ۷۰
شکل ۴-۶. خوشه‌بندی داده‌های اولیه در سه‌سطح متداخل توسط مدل 3-NCLUS ..... ۷۰

## فهرست جدول‌ها

**عنوان** **صفحه**


جدول ۲-۱. طبقه‌بندی الگوریتم‌های خوشه‌بندی..... ۱۳
جدول ۲-۲. توابع هدف به‌کار رفته در ساختار مدل‌های خوشه‌بندی ..... ۲۰
جدول ۲-۳. کاربرد فواصل مختلف بین داده‌ها (Rui, 2005) ..... ۲۸
جدول ۳-۱. نتایج حاصل از مجموع مربعات درون خوشه در مجموعه‌ی ۸۰تایی..... ۴۰
جدول ۳-۲. نتایج حاصل از مجموع مربعات درون خوشه در مجموعه‌ی ۹۰تایی..... ۴۲
جدول ۳-۳. نتایج حاصل از مجموع مربعات درون خوشه در مجموعه‌ی ۲۱۶ تایی ..... ۴۴
جدول ۳-۴. درصد خلوص خوشه‌بندی برای پایگاه داده‌ی Iris ..... ۴۶
جدول ۳-۵. درصد خلوص خوشه‌بندی برای پایگاه داده‌ی Original Owner and Donor..... ۴۷
جدول ۳-۶. درصد خلوص خوشه‌بندی برای پایگاه داده‌ی Glass ..... ۴۸
جدول ۳-۷. تحلیل حساسیت تعداد داده با حل خطی مدل BGSD..... ۵۰
جدول ۳-۸. تحلیل حساسیت تعداد خوشه با حل خطی مدل BGSD ..... ۵۱
جدول ۳-۹. تحلیل حساسیت تعداد داده با حل غیرخطی مدل BGSD..... ۵۱
جدول ۳-۱۰. تحلیل حساسیت تعداد خوشه با حل غیرخطی مدل BGSD ..... ۵۲
جدول ۴-۱. تعداد خوشه‌های هر دوسطح از خوشه‌بندی ..... ۶۴
جدول ۴-۲. تعداد خوشه‌های سه‌سطحی خوشه‌بندی..... ۶۹
جدول ۴-۳. توابع هدف تعیین شده در هر یک از سطوح خوشه‌بندی ..... ۷۳

## فهرست نمودارها

**عنوان** **صفحه**

نمودار ۳-۱. ارزیابی مجموع مربعات درون خوشه روی داده‌های مجازی ..... ۴۵
نمودار ۳-۳. درصد خلوص مدل‌های موجود بر روی پایگاه داده‌های واقعی ..... ۴۹
نمودار ۳-۴. تغییرات زمان در افزایش جداگانه‌ی یک واحد داده و یک خوشه در حل خطی مدل BGSD ..... ۵۳
نمودار ۳-۵. تغییرات زمان در افزایش جداگانه‌ی یک واحد داده و یک خوشه در حل غیرخطی مدل BGSD ..... ۵۳

# فصل اول: ورودی به پژوهش و شناخت آن

- ✓ **تعریف مسئله و بیان موضوع پژوهش**
- ✓ **سؤالات، اهداف و لزوم پژوهش**
- ✓ **اصطلاحات پایه**
- ✓ **ساختار کلی پایان‌نامه**

## ۱-۱ سرآغاز فصل نخست

در این فصل موضوع اصلی پژوهش، اهداف طرح مسئله و لزوم ارائه‌ی آن تشریح می گردد. سپس خلاصه‌ای از روند انجام کار توضیح داده می‌شود و اصطلاحات پایه‌ای نیز معرفی می شوند.

## ۱-۲ تعریف مسئله و بیان موضوع پژوهش

خوشه‌بندی، یکی از روش‌های طبقه‌بندی داده‌هاست. این روش، داده‌ها را بر اساس ماهیت ذاتی، در کنار یکدیگر قرار می‌دهد. ساده‌ترین روش خوشه‌بندی، از طریق مشاهده‌ی داده‌ها و نمایش دو بعدی آن‌هاست. از آن‌جا که داده‌ها در اکثر مواقع دارای ویژگی‌های متعددی هستند، تعداد خوشه‌های آن‌ها قابل تشخیص نیست. وجود این‌گونه نمونه‌ها در عالم واقعی باعث ایجاد مدل‌های مختلفی از خوشه‌بندی شده است. محققان رده‌بندی‌های مختلفی برای الگوریتم‌های خوشه‌بندی ارائه داده‌اند. مهمترین این رده‌بندی‌ها، گروه‌بندی سه‌تایی شامل مدل‌های خوشه‌بندی افرازی، مدل‌های خوشه‌بندی سلسله‌مراتبی و خوشه‌بندی فازی است و گروه‌بندی‌های دیگر با شکست ساختار سلسله‌مراتبی و افرازی به وجود آمده‌اند. خوشه‌بندی افرازی داده‌ها، شامل مدل‌هایی است که داده‌ها را در تعداد خوشه‌های از پیش تعیین شده قرار می‌دهد و خوشه‌بندی سلسله‌مراتبی نیز شامل مدل‌هایی است که به صورت پله‌ای و بدون نیاز به تعداد خوشه، داده‌ها را خوشه‌بندی می‌کند. الگوریتم‌های سلسله‌مراتبی، ساختار درختی دارند و به تعداد خوشه نیاز ندارند. به همین دلیل نسبت به خوشه‌بندی افرازی دارای برتری هستند. استفاده از الگوریتم‌های سلسله‌مراتبی به عنوان یک جواب اولیه برای خوشه‌بندی افرازی، اولین پل ارتباطی بین حوزه‌ی خوشه‌بندی افرازی و سلسله‌مراتبی است.

تاکنون محققان برای پاسخگویی به سؤالات چگونگی خوشه‌بندی خالص داده‌ها و تشخیص خوشه‌های مشابه و میزان نزدیکی آن‌ها به یکدیگر، بعد از خوشه‌بندی افرازی داده‌ها، به تحلیل خوشه‌ها پرداخته‌اند. مدل خوشه‌بندی

تک‌سطحی گام را فراتر می‌گذارد. به‌گونه‌ای که مدل خوشه‌بندی تودرتوی تک‌سطحی یک مدل خوشه‌بندی افرازی نوین و با رویکرد جداسازی خوشه‌ها از یکدیگر است. مدل خوشه‌بندی تودرتوی دوسطحی و بالاتر، بر پایه-ی مدل تک‌سطحی و با ماهیت سلسله‌مراتبی و توسط ابزار بهینه‌سازی، علاوه بر خوشه‌بندی داده‌ها، خوشه‌های شبیه به هم را نیز تعیین می‌کند. مدل خوشه‌بندی تودرتوی سه‌سطحی و بالاتر نیز می‌تواند علاوه بر خوشه‌بندی داده‌ها و تعیین خوشه‌های شبیه به هم، میزان نزدیکی خوشه‌ها به یکدیگر را نیز تشخیص می‌دهد.

## ۱-۳ سؤالات پژوهش

سؤالاتی که پژوهش بر اساس آن‌ها شکل گرفت:

- چگونه می‌توان خوشه‌بندی چند سطحی را با ابزار بهینه‌سازی انجام داد؟
- چگونه می‌توان داده‌ها را در چند سطح، توسط یک الگوریتم خوشه‌بندی نمود؟
- چگونه می‌توان خوشه‌های شبیه به هم را توسط یک مدل نشان داد؟
- چه معیاری برای شباهت دو خوشه می‌توان در نظر گرفت؟
- چگونه می‌توان خوشه‌بندی را در سطوح مختلف و بدون شکست در خوشه‌های سطوح دیگر انجام داد؟

سؤالاتی که در حین پژوهش با آن مواجه شدیم:

- مدل خوشه‌بندی تودرتو بر اساس کدام تابع هدف و یا کدام مدل خوشه‌بندی افرازی ساخته شود؟
- مدل ریاضی عدد صحیح آمیخته بر اساس چه اصلی می‌تواند بهینگی سطوح مختلف خوشه‌بندی را حفظ کند؟

## ۱-۴ اهداف پژوهش

- ایجاد یک مدل خوشه‌بندی که با حفظ یکپارچگی سطوح خوشه‌بندی، بتواند داده‌ها را در سطوح مختلف قرار دهد.
- تعیین خوشه‌های شبیه به هم، به منظور کشف الگوهای پنهان بین خوشه‌ها.
- استفاده از ابزار برنامه‌ریزی ریاضی در حوزه‌ی خوشه‌بندی تودرتو به عنوان ابزار بهینه‌یابی به جای روش-های ابتکاری.

## ۱-۵ ضرورت و لزوم پژوهش

با توجه به آن‌که تحلیل نتایج خوشه‌بندی از موارد بسیار مهم در مسیر کشف دانش است، وجود مدلی به منظور ساده کردن تحلیل نتایج، مفید به نظر می‌رسد. با توجه به آن‌که در بسیاری از موارد به منظور کشف الگوهای پنهانی در مجموعه‌ی داده‌ها، از روش‌های ابتکاری در خوشه‌بندی دوسطحی استفاده شده است، این تحقیق بر مبنای روش بهینه‌سازی و در سطوح پیوسته، خوشه‌بندی چندسطحی داده‌ها را انجام می‌دهد. در این مدل، خوشه‌های هر سطح به صورت پله‌ای وارد سطح دیگر شده و خوشه‌های شبیه به هم در سطوح بالاتر خوشه‌بندی، در کنار یکدیگر قرار می‌گیرند.

## ۱-۶ کاربردهای مدل

مدل خوشه‌بندی تودرتوی چندسطحی، در تمام موارد کاربرد خوشه‌بندی افرازی، مورد استفاده قرار می‌گیرد. از کاربردهای این مدل می‌توان حوزه‌ی بازاریابی، چیدمان، حوزه‌ی آموزشی، حوزه‌ی علوم زیستی و ژنتیک، روانشناسی و ... را نام برد. به طور مثال در مسائل بازاریابی می‌توان مثال زیر را مطرح نمود:

مدیریت فروشگاهی قصد دارد مشتریان خود را در ۱۰ گروه، خوشه‌بندی کند. علاوه بر آن می‌خواهد بداند کدام خوشه‌ها و چه‌میزان به یکدیگر شباهت دارد؟

مدل ارائه شده برای این مسئله 2-NCLUS است. این مدل، مشتریان را در سطح اول در۱۰ خوشه و سطح دوم۹ خوشه قرار می‌دهد. عدد۱۰ و۹ ورودی مسئله است. در نهایت مدل می‌تواند به صورت همزمان، دو سطح را خوشه‌بندی کند. در سطح اول کل مشتریان را در۱۰ خوشه و در سطح دوم، ۱۰ خوشه‌ی سطح اول را در۹ خوشه قرار می‌دهد و بدین صورت خوشه‌های نزدیک به هم در دو سطح تعیین می‌شوند.

## ۱-۷ مرور کلی بر روش‌شناسی پژوهش

قرار دادن داده‌ها در کنار یکدیگر بر اساس ماهیت خود داده‌ها و بدون افزودن برچسب خارجی بر روی آن‌ها خوشه‌بندی نام دارد. رویکردهای زیادی در حل مسایل خوشه‌بندی وجود دارد که بتواند داده‌ها را با بالاترین خلوص در خوشه‌ها قرار دهد. روش بهینه‌سازی یکی از آن روش‌ها است. بهینه‌سازی از روش‌های کارا در خوشه-بندی داده‌ها بوده و برنامه‌ریزی ریاضی نیز ابزار اجرای آن است.

در این پژوهش یک مدل ریاضی غیرخطی عدد صحیح آمیخته به منظور خوشه‌بندی چندسطحی متداخل داده‌ها، ارائه می‌شود. این مدل بر پایه‌ی خوشه‌بندی افرازی استوار است زیرا تعداد خوشه‌ها در هر سطح به مدل داده می‌شود و مدل به صورت یکپارچه داده‌ها را در تعداد سطوح خواسته شده خوشه‌بندی می‌کند. در ارائه‌ی این مدل خوشه‌بندی تودرتو، از یک مدل خوشه‌بندی افرازی که داده‌ها را با خلوص بالاتری خوشه‌بندی کند، استفاده می‌شود.

باتوجه به آن‌که فاصله‌ی بین داده‌ها اقلیدسی محاسبه می‌شود، داده‌های مورد استفاده عددی خواهند بود.

" در این تحقیق دو مدل ریاضی عدد صحیح آمیخته نیز در حوزه‌ی خوشه‌بندی افرازی ارائه شده است. مدل اول یک مدل غیرخطی ریاضی عدد صحیح آمیخته، با تابع هدف حداکثر کردن مجموع فواصل بین خوشه است که مدل BGSD نامیده می‌شود، این مدل با تکنیک تغییر متغیر به یک مدل خطی ریاضی عدد صحیح آمیخته تبدیل می‌شود. مدل دوم نیز یک مدل غیرخطی ریاضی عدد صحیح آمیخته، با تابع هدف حداقل کردن مجموع فواصل درون خوشه است که مدل WGSD نامیده می‌شود، این مدل تکمیل شده‌ی مدل ریاضی غیرخطی MINLP-SWGD است که (Saglam, 2005) در پروژه‌ی کارشناسی ارشد خود ارائه داده است. مدل WGSD نیز با تکنیک تغییر متغیر به یک مدل خطی ریاضی عدد صحیح آمیخته تبدیل می‌شود. به منظور بررسی خلوص مدل ریاضی پیشنهادی BGSD و مدل ریاضی تکمیل شده‌ی WGSD، از سه مجموعه پایگاه داده‌ی حقیقی و سه مجموعه پایگاه داده‌ی مجازی، استفاده و بر اساس آن اعتبار درونی و بیرونی مدل سنجیده می‌شود. به علاوه این مدل‌ها با مدل‌های ریاضی خوشه‌بندی افرازی دیگر محققان مقایسه و درجه‌ی اعتبار آن از طریق محاسبه‌ی معیار مجموع مربعات فواصل درون خوشه تعیین می‌شود. در همه‌ی مجموعه داده‌ها مدل پیشنهادی BGSD با درصد خلوص بالاتری خوشه‌بندی را انجام می‌دهد."

مدل خوشه‌بندی افرازی مورد استفاده در مدل‌سازی خوشه‌بندی تودرتو، مدل BGSD انتخاب می‌شود. بر این اساس تابع هدف آن در کلیه‌ی سطوح، حداکثر کردن مجموع فواصل بین خوشه‌ها است. این مدل خوشه‌بندی تودرتو که مدل K-NCLUS نامیده می‌شود، بر اساس اصل استقرا ارائه شده است. به طوری‌که در ابتدا مدل یک-سطحی، دوسطحی، سه‌سطحی ارائه و سپس بر پایه‌ی دو اصل *برقراری ارتباط بین سطح اضافه شده در خوشه-بندی و سطوح ماقبل و وجود پارامتری در تابع هدف به عنوان معیاری برای خوشه‌بندی سطح اضافه شده*، مسئله تا سطح $k$ام تعمیم داده می‌شود. برای اثبات درستی مدل ارائه شده نیز چند پایگاه داده‌ی مجازی مورد استفاده قرار می‌گیرد.

## ۸-۱ تعاریف و اصلاحات پایه

| | |
|---|---|
| *HC* | *خوشه‌بندی سلسله‌مراتبی* |
| *NHC* | *خوشه‌بندی غیر سلسله‌مراتبی (افرازی)* |
| *WGSS* | *مجموع مربعات فواصل درون خوشه‌ها* |
| *MWGD* | *حداقل کردن حداکثر فاصله در درون خوشه‌ها* |
| *TWGD* | *کلیه‌ی فواصل درون خوشه‌ها* |
| *SWGD* | *مجموع فواصل درون خوشه* |
| *WGAD* | *مجموع میانگین مربع فواصل در خوشه‌ها* |
| *MINLP* | *برنامه‌ریزی ریاضی غیرخطی عدد صحیح آمیخته* |
| *WGSD* | *مجموع فواصل درون خوشه* |
| *BGSD* | *مجموع فواصل بین خوشه* |
| *K-NCLUS* | *خوشه‌بندی تودرتو (متداخل) $k$ سطحی* |
| $R^m$ | *فضای $m$ بعدی* |
| $i$ | *داده‌ی $i$ ام* |
| $l$ | *خوشه‌ی $l$ ام در سطح اول از خوشه‌بندی* |
| $q$ | *خوشه‌ی $q$ ام در سطح دوم از خوشه‌بندی* |
| $h$ | *خوشه‌ی $h$ ام در سطح سوم از خوشه‌بندی* |
| *BIRCH* | *خوشه‌بندی سلسله‌مراتبی بر مبنای تعادل* |
| *CURE* | *خوشه‌بندی سلسله‌مراتبی بر اساس پایگاه داده‌ی بزرگ* |
| *GKA* | *خوشه‌بندی kmeans بر اساس الگوریتم ژنتیک* |
| *CHAMELEON* | *خوشه‌بندی بر اساس مدل‌های دینامیکی* |
| *ROCK* | *خوشه‌بندی بر اساس داده‌های طبقه‌ای* |
| *COOLCAT* | *خوشه‌بندی بر اساس سنجش آنتروپی روی داده‌های طبقه‌ای* |

## ۱-۹ ساختار کلی پایان‌نامه

ساختار کلی این تحقیق بدین شکل است که در فصل دوم پیشینه‌ی پژوهش در سه فاز بررسی می‌شود.

در فصل سوم دو مدل ریاضی در حوزه‌ی خوشه‌بندی افرازی ارائه می‌شود. سپس درصد خلوص و اعتبار مدل-های ارائه شده در مقایسه با مدل‌های محققان گذشته، از روی چند سری مجموعه داده‌های واقعی و مجازی بدست می‌آید و نتایج مورد تحلیل و بررسی قرار می‌گیرند. معیار قیاس مدل‌ها نیز مجموع مربعات درون خوشه است.

در فصل چهارم یک مدل ریاضی در خوشه‌بندی چند سطحی متداخل داده‌ها ارائه می‌شود. این مدل، نوین و حوزه‌ی جدیدی در خوشه‌بندی را به خود اختصاص می‌دهد. مدل خوشه‌بندی تودرتوی ارائه شده بر پایه‌ی مدل خوشه‌بندی افرازی که در فصل سوم مورد تأیید قرار گرفت، ساخته و با مدل ابتکاری محققان گذشته مقایسه می-شود. سپس مدل‌های خوشه‌بندی تودرتوی دیگری که در تابع هدف با یکدیگر تفاوت دارند، ارائه و با مدل خوشه-بندی اصلی ارزیابی می‌گردد.

در فصل ششم نیز نوآوری‌ها و دستاوردهای تحقیق بیان می‌شود و به سؤالات اساسی تحقیق پاسخ داده و نتایج کلی پژوهش ارائه می‌شود. سپس افق تحقیقات آتی تبیین می‌گردد.

## ۱-۱۰ سرانجام فصل نخست

در این فصل، مسئله‌ی اصلی پژوهش، روش پژوهش و لزوم انجام آن تشریح شد و ساختار کلی تحقیق مشخص گردید.

در فصل بعد ادبیات موضوع خوشه‌بندی در سه مرحله بیان می‌شود. دیدگاه محققان نسبت به مسئله‌ی خوشه-بندی و استفاده از ابزار حل برنامه‌ریزی ریاضی در حل مسایل خوشه‌بندی مطرح می‌گردد و جایگاه این پژوهش در میان تحقیقات گذشته تعیین می‌گردد.

# فصل دوم: بررسی پیشینه‌ی پژوهش

- ✓ طبقه‌بندی الگوریتم‌های خوشه‌بندی
- ✓ توابع هدف به کار رفته در مدل‌های خوشه‌بندی
- ✓ جایگاه موضوع پژوهش در تحقیقات گذشته
- ✓ سایر الگوریتم‌های خوشه‌بندی