
بسم الله الرحمن الرحیم

مرکز تهران

پایان‌نامه برای دریافت درجه کارشناسی ارشد

در رشته مهندسی کامپیوتر- نرم افزار

دانشکده فنی و مهندسی

گروه علمی مهندسی کامپیوتر و فناوری اطلاعات

عنوان پایان‌نامه:

ارائه روشی برای خوشه بندی و عملکرد آن روی داده‌های شرکت کنترل ترافیک تهران

نگارش:

الهه مرادی

استاد راهنما:

دکتر محمد مهدی عباد زاده

استاد مشاور:

دکتر احمد فراهی

زمستان ۱۳۹۰

جمهوری اسلامی ایران
وزارت علوم تحقیقات و فناوری

مرکز شمیرانات

تاریخ .....
شماره .....
پیوست .....

دانشگاه پیام نور

دانشگاه پیام نور استان تهران

تصویب نامه

پایان نامه کارشناسی ارشد رشته مهندسی کامپیوتر(نرم افزار)

تحت عنوان:

"ارائه روشی برای خوشه بندی و عملکرد آن روی داده های شرکت کنترل ترافیک تهران "

تاریخ دفاع :۱۳۹۰/۱۱/۳۰ ساعت: ۱۲-۱۰/۳۰

نمره:.....۱۸/۰..... درجه ارزشیابی: .عالی.....

هیات داوران:

داوران	نام و نام خانوادگی	مرتبه علمی	امضاء
استاد راهنما	دکتر محمد مهدی عبادزاده	دانشیار	
استاد مشاور	دکتر احمد فراهی	استادیار	
استاد داور	دکتر مهدی جوانمرد	استادیار	
نماینده گروه	دکتر محمد هادی معظم		

تهران- بزرگراه ارتش-انتهای
بلوار شهید مژدی (اوشان)
خیابان شهید پیروز شفیعی
خیابان یاران-خیابان یاران دوم
دانشگاه پیام نور مرکز شمیرانات
تلفن:٤-۲۲۱۹۵۳۰۳
دورنگار:۲۲٤۸٤۸۳٤
www.shemiranat.tpnu.ac.ir
shemiranat@tpnu.ac.ir

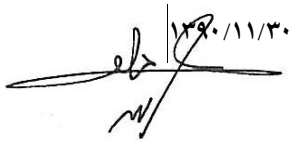
گواهی اصالت نشر و حقوق مادی و معنوی اثر

اینجانب **الهه مرادی** دانشجوی ورودی سال **۱۳۸٦** مقطع **کارشناسی ارشد** رشته **مهندسی کامپیوتر گرایش نرم افزار** گواهی می‌نمایم چنانچه در پایان نامه‌ی خود از فکر، ایده و نوشته دیگری بهره گرفته‌ام با نقل قول مستقیم یا غیر مستقیم، منبع و ماخذ آن را نیز در جای مناسب ذکر کرده‌ام. بدیهی است مسئولیت تمامی مطالبی که نقل قول دیگران نباشد؛ برعهده خویش می‌دانم و جوابگوی آن خواهم بود.

دانشجو تایید می‌نماید که مطالب مندرج در این پایان نامه، نتیجه تحقیقات خودش می‌باشد و در صورت استفاده از نتایج دیگران، مراجع آن را ذکر نموده است.

نام و نام خانوادگی دانشجو: **الهه مرادی**

تاریخ و امضاء ۱۳۹۰/۱۱/۳۰



اینجانب **الهه مرادی** دانشجوی ورودی سال **۱۳۸٦** مقطع **کارشناسی ارشد** رشته **مهندسی کامپیوتر گرایش نرم افزار** گواهی می‌نمایم چنانچه براساس مطالب پایان نامه‌ی خود، اقدام به انتشار مقاله، کتاب و ... به صورت مشترک و با ذکر نام استاد راهنما مبادرت نمایم.

نام و نام خانوادگی دانشجو: **الهه مرادی**

تاریخ و امضاء ۱۳۹۰/۱۱/۳۰

کلیه حقوق مادی مترتب از نتایج مطالعات، آزمایشات و نوآوری ناشی از تحقیق موضوع این پایان نامه متعلق به دانشگاه پیام نور می‌باشد.

بهمن ۱۳۹۰

به پاس تعبیر عظیم و انسانی‌شان از کلمه ایثار و ازخودگذشتگی،

به پاس عاطفه سرشار و گرمای امیدبخش وجودشان که در این سردترین روزگاران بهترین پشتیبان است،

به پاس قلب‌های بزرگشان که فریادرس است و سرگردانی و ترس در پناهشان به شجاعت می‌گراید،

و به پاس محبت‌های بی دریغشان که هرگز فروکش نمی‌کند،

این مجموعه را به **پدر و مادر** عزیزم تقدیم می‌کنم.

تشکر و قدردانی

« خداوندا به ما توفیق تلاش در شکست، صبر در نومیدی، رفتن بی‌همراه، جهاد بی-سلاح، کار بی‌پاداش، فداکاری در سکوت، دین بی‌دنیا، مذهب بی‌عوام، عظمت بی‌نام، خدمت بی‌نان، ایمان بی‌ریا، خوبی بی‌نمود، گستاخی بی‌خامی، مناعت بی‌غرور، عشق بی‌هوس، تنهایی در انبوه جمعیت و دوست داشتن بی آنکه دوست بداند، را عنایت فرما» (دکتر شریعتی)

سپاس خدایی را که هر گاه از او چیزی خواستیم عطا می‌فرماید و آنگاه که امیدی به او داشتیم به امیدمان می‌رساند. بر خود واجب می‌دانم که از زحمات بی‌دریغ، تلاش بی‌وقفه و راهنمایی‌های عالمانه استاد ارجمند جناب آقای دکتر عبادزاده در طول این پروژه تشکر و قدردانی به عمل آورم. همچنین از زحمات استاد محترم جناب آقای دکتر فراهی که با راهنمایی‌های خود رهگشای اینجانب بوده‌اند، تشکر می‌نمایم. امیدوارم بتوانم قدردان زحمات تمام عزیزانی باشم که مرا در این پروژه یاری کرده‌اند.

الهه مرادی

چکیده

ما در عصری قرار داریم که اغلب تحت عنوان عصر اطلاعات شناخته می‌شود. در این عصر، سازمان‌ها به وسعت در حال جمع آوری حجم بسیار عظیمی از اطلاعات هستند و به استخراج دانش نیازمند هستند. داده کاوی حاصل تکامل پنجاه ساله فناوری اطلاعات است. در دهه شصت میلادی اولین گام در جهت داده کاوی با جمع آوری داده‌ها برداشته شد، در دهه هفتاد سیستم‌های مدیریت پایگاه داده رابطه‌ای توسعه یافتند و در دهه هشتاد تکنیک‌های دست‌یابی به داده‌ها ارتقا پیدا کردند. در دهه نود زمینه اصلی داده کاوی با توسعه انبارهای داده و سیستم‌های پشتیبانی تصمیم گیری فراهم گردید. به وضوح، داده کاوی وقتی ظهور یافت که حجم اطلاعات جمع آوری شده از کمیتی که یک کاربر انسانی می‌تواند تفسیر کند فراتر رفت. خوشه بندی یکی از تکنیک‌های داده کاوی است که نقش مهمی در تحلیل و تفسیر داده‌ها ایفا می‌کند، خوشه بندی داده‌ها را در خوشه‌ها به گونه‌ای گروه بندی می‌کند که اشیاء داده‌ای درون یک خوشه به یکدیگر شباهت بسیار زیادی دارند، اما با اشیاء درون خوشه‌های دیگر شباهت بسیار کمی دارند. تحقیقات بسیار زیادی در توسعه روش‌های خوشه بندی فازی انجام گرفته است، که مورد استفاده ترین آنها الگوریتم‌های مبتنی بر تابع هدف هستند، که داده‌ها را توسط مینیمم کردن یک تابع هدف خوشه بندی می‌کنند. در بین الگوریتم‌های خوشه بندی فازی مبتنی بر تابع هدف، الگوریتم Fuzzy C-Means تاکنون معروف‌ترین الگوریتم شناخته شده است و مبنای خوبی برای بسیاری از روش‌های جدید مبتنی بر تابع هدف می‌باشد. گرچه، Fuzzy C-Means بسیار به داده‌ها حساس است و فقط وقتی می‌تواند خوب عمل کند که خوشه‌ها تقریباً اندازه و شکل کروی مشابهی داشته باشند. در این تحقیق سعی بر آن شده است که به دو مسأله اساسی، یکی انتخاب نامناسب مراکز اولیه و در نتیجه گیر افتادن در مینیمم محلی و دیگری پیدا کردن خوشه‌هایی با شکل و اندازه یکسان در خوشه بندی فازی پرداخته شود. با ارائه یک مرحله پیش پردازشی، قابل اعمال بر الگوریتم‌های فازی موجود و روش جدید ارائه شده، از حساسیت الگوریتم به داده‌ها کاسته می‌شود و مراکز خوشه به نوعی انتخاب می‌شوند که از گیر افتادن در مینیمم محلی اجتناب می‌گردد. سپس تابع فاصله‌ای با کمک مقادیر ویژه و بردارهای ویژه ماتریس کواریانس فازی خوشه‌ها برای خوشه بندی ارائه می‌گردد که در شناسایی خوشه‌های به خوبی مجزا کمک شایان توجهی می‌کند. بنا بر معیارهای ارزیابی روش جدید برتری خود را نسبت به الگوریتم‌های موجود اثبات می‌کند. در ادامه به علت اهمیتی که مسأله ترافیک در جوامع امروزی ایفا می‌کند با استفاده از روش پیشنهادی روی داده‌های شرکت کنترل ترافیک تهران داده کاوی انجام شده است و نتایج آن ارائه گردیده‌اند.

کلمات کلیدی

داده کاوی، خوشه بندی، خوشه بندی فازی، درجه عضویت، تابع هدف، تابع فاصله، معیار ارزیابی، اعتبارسنجی خوشه، شاخص اعتبار.

فهرست مطالب

فهرست جداول

فهرست شکل‌ها

فصل اول

مقدمه

کامپیوترهای اولیه اساساً، برای سر و کار داشتن با اعداد طراحی شده بودند. وقتی حافظه قابل تهیه‌تر شد، شروع به جمع آوری داده‌ها[1] با نرخ افزایشی کردیم. دست‌کاری داده‌ها، اطلاعات[2] را به واسطه گونه‌های شگفت آور سیستم‌ها و کاربردهای هوشمند ایجاد کرد. وقتی جمع آوری داده ادامه یافت و اطلاعات بیشتری حاصل شد، سطح دیگری از تقطیر برای تولید دانش[3] اضافه شد. دانش اساس اطلاعات است و طعم‌های مختلفی دارد. سیستم‌های خبره[4]، پایگاه‌های دانش[5]، سیستم‌های پشتیبانی تصمیم گیری[6]، یادگیری ماشینی[7]، سیستم‌های خود مختار[8]، و عامل‌های هوشمند[9] برخی از بسته‌های تحقیقاتی متعددی هستند، که به منظور توصیف کاربردهایی که بخشی از قابلیت‌های ذهنی انسان را تقلید می‌کنند ابداع شده‌اند. فرآیند بسیار موفق و وسیعاً محبوب استخراج دانش از کوهی از داده‌ها داده کاوی[10] نیز از این دست می‌باشد.

۱–۱ تعریف مسأله و سوالات اصلی تحقیق

این تحقیق به ارائه روشی برای خوشه بندی داده‌ها اختصاص یافته است. خوشه بندی یکی از تکنیک‌های داده کاوی است و به شناسایی تعداد متناهی خوشه که داده‌ها را بر اساس شباهت‌ها و تفاوت‌هایشان گروه بندی می‌کند می‌پردازد. الگوریتم‌های متعددی وجود دارند که به خوشه بندی داده‌ها

[1] Data
[2] Information
[3] Knowledge
[4] Expert Systems
[5] Knowledge Bases
[6] Decision Support Systems
[7] Machine Learning
[8] Autonomous Systems
[9] Intelligent Agents
[10] Data Mining

می‌پردازند، هر یک از این الگوریتم‌ها دارای مزایا و معایبی هستند (Pavel B., 2002). در این میان تحقیقات بی‌شماری در زمینه توسعه روش‌های خوشه بندی فازی انجام گرفته است، که پر استفاده ترین آنها الگوریتم‌های مبتنی بر تابع هدف هستند، که داده‌ها را توسط مینیمم کردن یک تابع هدف خوشه بندی می‌کنند. در بین الگوریتم‌های مبتنی بر تابع هدف الگوریتم Fuzzy C-Means معروف‌ترین الگوریتم است و مبنای متداولی برای اکثر روش‌های مبتنی بر تابع هدف جدید می‌باشد (Zhang H., 2005). این الگوریتم با وجود کاربرد بسیاری که دارد ، در شناسایی مناسب مراکز اولیه برای خوشه‌ها ناتوان است و گاهی این امر موجب گیر افتادن آن در مینیمم محلی می‌شود، در ضمن خوشه‌ها در این الگوریتم فقط شکل کروی دارند. بنابراین در این پایان نامه الگوریتم‌های مختلف خوشه بندی بررسی شده‌اند و سعی بر آن است که روشی در خوشه بندی داده‌ها ابداع گردد. هدف این است که روش پیشنهادی تا بیشترین حد ممکن از نواقص الگوریتم فازی موجود از جمله ناتوانی در تشخیص خوشه‌های مفید و مشکلات مربوط به شکل و اندازه خوشه‌ها و غیره مبرا باشد.

در ادامه به مسأله ترافیک پرداخته شده است. در این بخش از داده‌های سیستم ثبت وقایع شرکت کنترل ترافیک تهران استفاده شده است. شرکت کنترل ترافیک تهران زیر مجموعه معاونت حمل و نقل و ترافیک شهرداری تهران می‌باشد و در جهت رفع مشکلات ترافیکی شهر تهران فعالیت می‌کند. استفاده از تکنولوژی‌های روز مخابرات، الکترونیک و کامپیوتر و کنترل پیشرفته ترافیک به منظور بهره‌وری کامل از پتانسیل‌ها و ظرفیت‌های ناوگان و شبکه از اهداف این شرکت در راستای حل مشکلات ترافیکی است. داده کاوی یک تکنولوژی کامپیوتری است که اطلاعات با ارزش و پنهان را از پایگاه داده‌ها کشف می‌کند و از جمله اموری است که هر کشور، سازمان و شرکتی به منظور توسعه علمی، فنی و اقتصادی خود به آن نیاز دارد (Haluzov P., 2008). از این رو داده کاوی به عنوان یکی از تکنولوژی‌های جدید روز می‌تواند جهت کشف الگوهای مناسب ترافیکی بسیار مثمر ثمر باشد. بنابراین در ادامه سعی می‌شود بر روی اطلاعات سیستم ثبت وقایع شرکت کنترل ترافیک تهران با استفاده از روش پیشنهادی برای خوشه بندی به داده کاوی پرداخته شود. با انجام این مرحله دو هدف اصلی تحقیق که یکی ارائه یک الگوریتم برای خوشه بندی داده‌ها و بررسی توان عملیاتی الگوریتم ارائه شده و دیگری شناسایی یک الگوی مفید ترافیکی است مورد ارزیابی قرار می‌گیرند.

در واقع در پایان این تحقیق به سوالات زیر بایستی پاسخ داد:

۱- آیا روش پیشنهادی برای خوشه بندی داده‌ها از نقایص الگوریتم‌های موجود مبرا است؟

۲- آیا روش پیشنهادی برای خوشه بندی داده‌ها کارایی مناسبی دارد؟

۳- نتیجه داده کاوی روی اطلاعات سیستم ثبت وقایع شرکت کنترل ترافیک تهران چیست؟

۱-۲ سابقه و ضرورت انجام تحقیق

دامنه کاربرد داده کاوی و تکنیک‌ها و تکنولوژی‌های مرتبط با آن در پنج سال اخیر بسیار توسعه یافته‌اند. توسعه ابزارهای جمع آوری داده اتوماتیک و به دنبال آن انفجار داده‌ای مهیب، نیاز حتمی به تفسیر و بهره‌برداری بهتر از حجم عظیم داده‌ها را تقویت کرد. بهبود مداوم سخت افزار به همراه وجود الگوریتم‌های پشتیبان، توسعه و پیشرفت متدلوژی‌های داده کاوی پیچیده را ممکن ساخته‌اند. روش‌های متعددی برای تحقق ابزارهای داده کاوی خود مختار و مستعد به جهت پشتیبانی از همه گام‌های قبل و بعد از فرآیند کشف دانش در پایگاه داده‌ها اقتباس شده‌اند (Fayyad U. M., et al., 1996).

دو دلیل اساسی برای استفاده از داده کاوی وجود دارد: ۱) وجود حجم انبوهی از داده‌ها و در عین حال فقدان اطلاعات کافی ۲) نیاز به استخراج اطلاعات از درون داده‌ها و تفسیر این داده‌ها.

در روبرویی با حجم عظیم داده‌ها، تحلیلگران انسانی بدون ابزارهای خاص، دیگر به اندازه کافی کارایی ندارند. درحالیکه داده کاوی می‌تواند فرآیند پیدا کردن ارتباطات و الگوها از داده‌های خام را اتوماتیک کند، و نتایج آن می‌توانند یا در سیستم‌های پشتیبانی تصمیم گیری اتوماتیک مورد استفاده قرار گیرند یا در اختیار یک تحلیلگر قرار داده شوند. به این دلیل است که از داده کاوی، به خصوص در حوزه علم و تجارت که نیاز به تحلیل حجم زیادی از داده‌ها برای کشف رویه‌ها وجود دارد استفاده می‌شود.

به داده کاوی می‌توان به عنوان وارث منطقی تکنولوژی اطلاعات اندیشید (Bigus J.P., 1996). با نظر به تکامل IT در ۵۰ سال گذشته (Pilot Software Inc., 1999)، می‌بینیم که اولین گام اساسی در دهه شصت با جمع آوری داده‌ها برداشته شد، در حالیکه در دهه هفتاد، اولین سیستم مدیریت پایگاه داده رابطه‌ای[1] توسعه داده شد. در طول دهه هشتاد، تکنیک‌های ارتقا یافته دست‌یابی به داده‌ها شروع به پدیدار شدن کردند، مدل رابطه‌ای به وسعت به کار گرفته شد، و زبان‌های برنامه نویسی مناسب توسعه داده شدند (Bigus J.P., 1996). به زودی در دهه نود، گام معنی دار دیگری در مدیریت داده‌ها برداشته

[1] Relational Database Management System

شد. توسعه انبارهای داده[1] و سیستم‌های پشتیبانی تصمیم گیری اجازه دست‌کاری داده‌های به دست آمده از منابع ناهمگن و پشتیبانی از پویایی چند سطحی و خلاصه سازی داده‌ها را دادند.

گرچه پیشرفت‌های ایجاد شده توسط سیستم‌های پشتیبانی تصمیم گیری و کارایی انبارهای داده برانگیزنده هستند، اما به تنهایی نمی‌توانند راه حل راضی کننده‌ای برای مسأله غنی بودن داده اما فقدان دانش، که به ابزارهای تحلیل داده پیشرفته نیاز دارد، فراهم کنند (Information Discovery Inc., 1999). تلاش انسان برای دانش و ناتوانی در درک حجم داده‌های- دایماً افزایش یابنده- یک سیستم منجر به چیزی که امروزه آن را داده کاوی می‌خوانیم شد. به وضوح، داده کاوی وقتی ظهور یافت که حجم اطلاعات جمع آوری شده از کمیتی که یک کاربر می‌تواند تفسیر کند فراتر رفت.

روش‌های داده کاوی متداول شامل دسته بندی[2]، پس گرایی[3]، خوشه بندی[4]، تلخیص[5]، مدل کردن وابستگی[6]، و تشخیص تغییر و انحراف[7] می‌باشند (Fayyad U.M, et al., 1996). خوشه بندی یکی از روش‌های مهم در داده کاوی است و در بردارنده شناسایی یک مجموعه محدود از خوشه‌ها[8] برای توصیف داده‌ها است. خوشه‌ها می‌توانند متقابلاً منحصر به فرد[9]، سلسله مراتبی[10] یا همپوشان[11] باشند (Fayyad U.M, et al., 1996). هر عضو از یک خوشه باید بسیار شبیه به بقیه اعضای آن خوشه و بسیار بدون شباهت با اعضای خوشه‌های دیگر باشد (Han J., Kamber M., 2001).

خوشه بندی دارای الگوریتم‌های متعددی است و تحقیقات بسیار زیادی در جهت توسعه و اصلاح این الگوریتم‌ها صورت گرفته است (Pavel B., 2002). آنچه امروزه و البته در این تحقیق بیشتر مد نظر است بررسی و اصلاح الگوریتم‌های خوشه بندی می‌باشد، ضرورت انجام چنین کاری برطرف کردن بخشی از نواقص برخی از الگوریتم‌های خوشه بندی موجود و در واقع ارائه روشی است که بتواند به بهترین شکل ممکن داده‌ها را خوشه بندی کند.

---

[1] Data Warehouse
[2] Classification
[3] Regression
[4] Clustering
[5] Summarization
[6] Dependency Modeling
[7] Change and Deviation Detection
[8] Clusters
[9] Mutually Exclusive
[10] Hierarchical
[11] Overlapping

داده کاوی روی داده‌های سیستم‌های کنترل ترافیک امروزه به بخشی از برنامه‌های اصلی کشورهای پیشرفته تبدیل شده است. هدف از انجام داده کاوی در اکثر کارهای انجام شده مدیریت ترافیک با شناسایی مکان‌های پر تردد، مناطق تصادف خیز، شناسایی عوامل منجر به حجم بالای ترافیک و تصادفات، کشف الگوهای ترافیکی و غیره می‌باشد. در پنج سال اخیر پروژه‌های داده کاوی زیادی در زمینه ترافیک در کشورهای مختلف با مقیاس‌های گوناگون در سطوح متفاوت منطقه‌ای، شهری، استانی و... صورت گرفته است (Haluzov P., 2008; Bladi M., et al., 2006; Chang T.L., et al., 2005). در تهران در این زمینه تلاش‌هایی صورت گرفته، اما بنا به گزارش شرکت کنترل ترافیک تهران تا کنون کار داده کاوی جامعی با نتیجه مشخص ارائه نشده است. بنابراین با توجه به اوضاع ترافیک تهران و نیاز مبرم شهروندان به حل معضل ترافیک و نقش موثری که داده کاوی می‌تواند در شناسایی این معضل داشته باشد، ضرورت انجام یک پروژه داده کاوی کاملاً محسوس است.

۱–۳ فرضیه‌ها

- فرض بر این است که با انجام مرحله پیش پردازش، الگوریتم پیشنهادی در مینیمم محلی گیر نیفتند و همواره پاسخ‌های بهینه سراسری تولید کند.
- فرض بر این است که با ارائه یک تابع فاصله جدید خوشه‌هایی بیضی شکل و مجزاتر در تعداد تکرارهای کمتر الگوریتم تولید شود. و بدین ترتیب بر بخشی از نقایص الگوریتم خوشه بندی فازی موجود غلبه کرده و کارایی مناسبی ارائه شود.
- فرض بر این است که داده کاوی روی داده‌های شرکت کنترل ترافیک تهران به شناسایی برخی از مشکلات ترافیکی این شهر کمک کند.

۱–۴ اهداف تحقیق

- ارائه روشی برای خوشه بندی داده‌ها که از بخشی از نواقص برخی از الگوریتم‌های موجود از جمله ناتوانی در تشخیص مراکز خوشه‌ی مفید و مشکلات مربوط به شکل و اندازه خوشه‌ها مبرا باشد.
- داده کاوی روی داده‌های شرکت کنترل ترافیک تهران با استفاده از روش پیشنهادی و کمک به شناسایی برخی از مشکلات ترافیکی.