دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

پایان‌نامه‌ی کارشناسی

عنوان:

دسته‌بندی خانوارهای ایرانی براساس وضعیت رفاهی ـ اقتصادی و ارائه معیاری برای سنجش کیفیت دهک‌بندی

نگارش:

محمدصدرا حیدری ـ علیرضا ایلامی

استاد راهنما:

جناب آقای دکتر محمدامین فضلی

شهریور ۱۴۰۱

فصل ۱

دادگان

۱ـ۱ معرفی داده

در این پروژه از داده‌های عمومی «پایگاه اطلاعات رفاه ایرانیان» استفاده می‌کنیم. این پایگاه اطلاعات نتیجه تلاش وزارت تعاون، کار و رفاه اجتماعی جمهوری اسلامی ایران برای ایجاد شناسنامه رفاهی‌ـ اقتصادی افراد است.

پایگاه اطلاعات رفاه ایرانیان، از تجمیع ۵۰ منبع داده‌ای تشکیل شده‌است. در فاز نخست، ۲۵ منبع داده‌ای به طور کامل تجمیع و ساختار اصلی پایگاه اطلاعات رفاه ایرانیان (با بیش از ۶۰ جدول داده‌ای) بر اساس آن‌ها شکل گرفته‌است. بیش از ۳ میلیارد رکورد داده‌ای در پایگاه داده نهایی ذخیره و تجمیع شده است. در حال حاضر، ۲۲۱ فیلد اطلاعاتی شناسایی شده‌اند که ابعاد مختلف هویتی شهروندان را به صورت مستقیم و غیرمستقیم توصیف می‌نمایند. بخشی از خروجی‌های پایگاه اطلاعات رفاه ایرانیان به صورت وب‌سرویس در اختیار ۲۰ سازمان مختلف قرار گرفته است. علاوه‌برآن، خدمات تحلیلی نیز به تعدادی از سازمان‌ها ارائه شده است. این خدمات شامل استحقاق‌سنجی، بررسی همپوشانی، داده‌کاوی، وسع‌سنجی و ... می‌شود. (لینک به صفحه)

پایگاه ملی اطلاعات رفاه ایرانیان حاوی اطلاعات مفیدی در مورد ابعاد مختلف اجتماعی‌ـ اقتصادی شهروندان ایرانی است که می‌تواند برای اهداف مختلف مورد تحلیل و بررسی اندیشمندان، جامعه‌شناسان، اقتصاددانان و به طور کلی جامعه محققان کشور قرار گیرد. از سمتی دیگر، این پایگاه شامل اطلاعات

محرمانه و خصوصی شهروندان است که نقض این حریم خصوصی، ممکن است تبعات جبران ناپذیری در زمینه اعتماد عمومی ایجاد نماید. معاونت رفاه اجتماعی به منظور ایجاد یک مصالحه میان ۱ـ امکان دسترسی جامعه محققان به داده‌های پایگاه اطلاعات ایرانیان جهت انجام مطالعات تحقیقاتی و ۲ـ حفظ حداکثری حریم خصوصی شهروندان ایرانی، اقدام به تهیه نمونه‌های ۲ درصدی از اطلاعات موجود در پایگاه ملی اطلاعات رفاه ایرانیان نموده است. در این نمونه‌برداری‌ها سعی شده است که نکات زیر به صورت جدی مدنظر قرار گیرند.

۱ـ۲ معرفی ویژگی‌ها

دیتاست موجود شامل اطلاعات پانصدهزار خانوار و ۱/۴۹۰/۹۹۱ فرد ایرانی است و برای هر فرد ۴۸ ویژگی دارد. ویژگی‌های موجود در دیتاست را می‌توان به دسته‌های زیر تقسیم‌بندی کرد.

- **اطلاعات شخصی**
  - شناسه فرد
  - تاریخ تولد
  - شناسه سرپرست خانوار
  - جنسیت
- **محل زندگی**
  - کد پستی
  - شهرستان محل زندگی
  - استان محل زندگی
  - شهری یا روستایی بودن
- **اطلاعات حساب** (برای سال‌های ۹۵ تا ۹۸ و به تفکیک سال)
  - گردش بستانکار
  - مانده ابتدای سال
  - گردش بدهکار
  - مانده انتهای سال
  - مجموع سود حساب‌ها
- **اطلاعات تراکنش** (سال ۹۸ و شش ماههٔ اول سال ۹۹ و به تفکیک سال)

- مقدار کل تراکنش کارت‌ها — تعداد کل تراکنش کارت‌ها

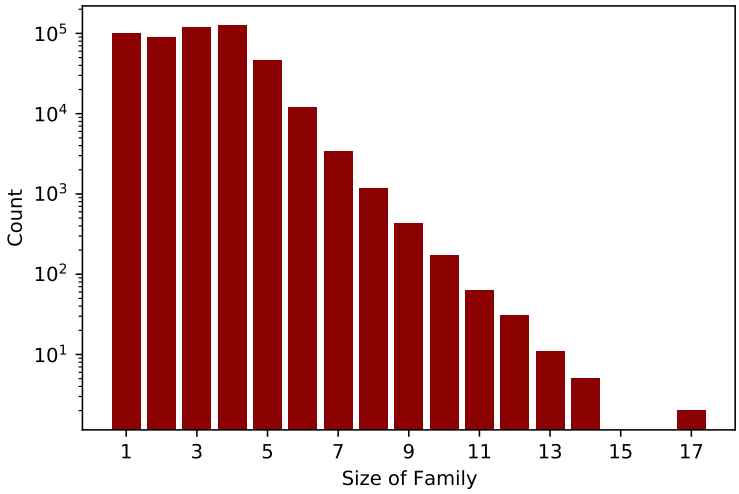
- **دارایی‌ها**
  - تعداد خودروهای فرد — مجموع ارزش خودروهای فرد
- **تفریحات** (در بازهٔ سال‌های ۹۶ تا ۹۹)
  - تعداد سفرهای خارجی هوایی غیرزیارتی — تعداد سفرهای خارجی هوایی زیارتی
  - تعداد سفرهای خارجی زمینی غیرزیارتی — تعداد سفرهای خارجی زمینی زیارتی
- **کسب و کار**
  - داشتن مجوز صنفی — صنفی که در آن مجوز دارد
- **بیمه**
  - داشتن بیمه سلامت — نوع بیمه سلامت
- **شرایط خاص**
  - داشتن بیماری خاص — معلول بودن فرد
- **بازنشستگی**
  - بیمه‌پرداز صندوق‌های بازنشستگی بودن — بازنشسته صندوق‌های بازنشستگی بودن
- **مالیات و درآمد**
  - شاغل مشمول مالیات بودن — مجموع درآمد فرد از حقوق

۱ـ۳ آماره‌ها و شهود

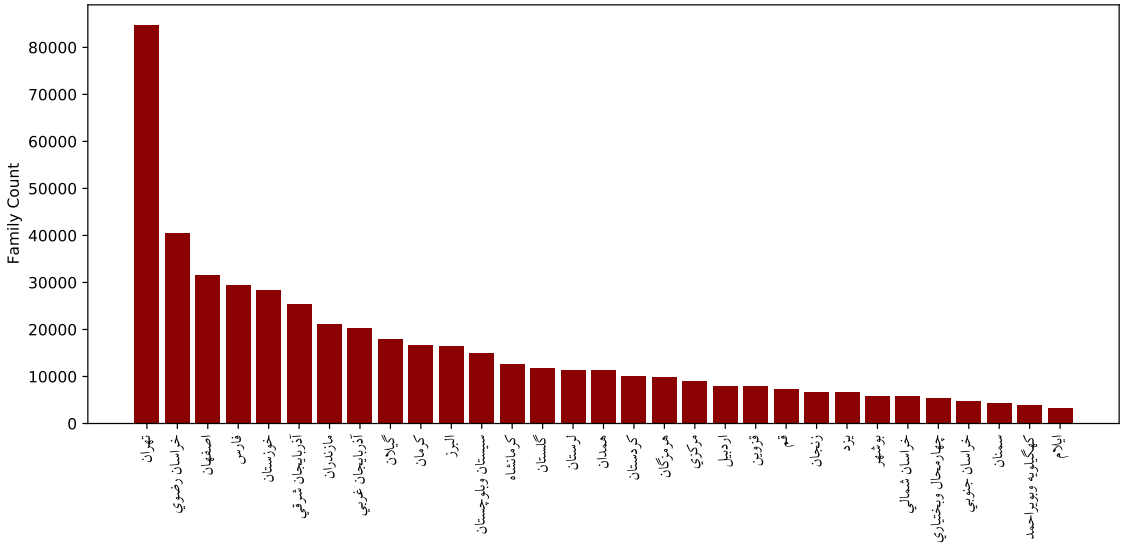
۱ـ۳ـ۱ اندازه خانوار

۱ـ۳ـ۲ توزیع خانوارها در کشور

۱ـ۳ـ۳



شکل ۱ــ۱: نمودار هیستوگرام اندازهٔ خانوارهای موجود در نمونه ۲ درصدی پایگاه اطلاعات رفاه ایرانیان. در نمودار بالا محور عمودی به صورت لگاریتمی مقیاس شده‌است.



شکل ۱ــ۲: نمودار تعداد خانوارهای موجود در نمونه ۲ درصدی پایگاه اطلاعات رفاه ایرانیان به تفکیک استان محل زندگی. بیشترین تعداد خانوار مربوط به استان تهران و کمترین آن مربوط به استان ایلام می‌باشد.

فصل ۲

روش‌شناسی

۲ـ۱ روند کلی

برای دسته‌بندی خانوارها به کمک روش‌های مبتنی بر هوش مصنوعی، لازم است هر خانوار را در فضای برداری نمایش دهیم. استفاده مستقیم از ویژگی‌های موجود در دیتاست رفاه ایده مناسبی نیست چرا که اهمیت هر ویژگی با دیگری در تقسیم‌بندی ما تفاوت دارد.

برای حل این موضوع ابتدا ویژگی‌ها را به چند دسته تقسیم می‌کنیم. سپس برای هر دسته معیار شباهتی مشخص می‌کنیم. این معیار شباهت باید به گونه‌ای باشد که با دریافت ویژگی‌های مشخص از دو خانوار، میزان شباهت آن‌ها را در مقیاس ۰ (کاملاً متفاوت) تا ۱ (کاملاً یکسان) خروجی دهد. سپس به‌ازای هر دسته از ویژگی‌ها یک ماتریس شباهت تولیدکرده که درایه ij آن نشان‌دهنده میزان شباهت خانوار iام و jام خواهد بود. بدیهی‌است که داده‌های قطری در این ماتریس همگی برابر با ۱ می‌باشند.

در ادامه با استفاده از الگوریتم t-SNE ماتریس‌های ساخته‌شده را به فضای برداری با بعد دلخواه n نگاشت می‌کنیم به صورتی که فاصله اقلیدسی خانوارهایی که در ماتریس شباهت زیادی باهم داشته‌اند کم بوده و فاصله خانوارهای بی‌شباهت زیاد باشد. این عمل را برای همه دسته‌ها انجام داده و با به‌هم چسباندین نتایج حاصل هر خانوار را در فضای برداری با بعد دلخواه نگاشت کرده‌ایم.

بعد از آن تلاش می‌کنیم با استفاده از الگوریتم‌های مختلف دسته‌بندی، خانوارها را در فضای برداری ساخته‌شده دسته‌بندی کرده و تا حد امکان آن‌ها را از یکدیگر تفکیک کنیم.

باید توجه کرد که به واسطه استفاده از الگوریتم t-SNE نتایج حاصل مقداری تصادفی بوده ولی در هدف و نتیجه کار ما تأثیری ایجاد نمی‌کند.

Sharif University of Technology

Department of Computer Engineering

M.Sc. Thesis

Iranian Household Clustering and developing new metric to test qaulity of deciles

By:

Mohammad Sadra Heydari - Alireza Ilami

Supervisor:

Dr. Mohammad Amin Fazli

September 2020