به نام خدا

دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر

گزارش پروژه نهایی درس پردازش داده‌های حجیم

استاد درس: دکتر مصطفی حقیر چهرقانی

اعضای گروه:

زهرا سادات رضوی نژاد - (۹۹۱۳۱۰۴۴)

فاطمه غلام‌زاده - (۹۹۱۳۱۰۰۳)

نیم سال دوم ۱۴۰۰-۱۳۹۹


بسم الله الرحمن الرحيم

فهرست مطالب **صفحه**

فهرست اشکال و جداول **صفحه**

موضوع مورد بحث

در این پژوهش یک روش دسته‌بندی برمبنای گراف، به منظور یادگیری نیمه‌نظارت شده[1] برای داده‌های اقلیدسی و دسته‌بندی داده‌های گرافی پیشنهاد شده‌است. با تغییرات اعمال‌شده روی تابع هزینه سعی شده الگوریتم مورد نظر را نسبت به نویز مقاوم کند.

در اینجا تعدادی از نقاط برچسب دارند و تعدادی از آنها برچسب ندارند و هدف پیش‌بینی برچسب داده‌های نوع دوم است. داده‌ها هم در یک فضای با ابعاد بالا یا با یک ساختار گرافی فرض شده اند. در واقع یک مسئله‌ی بهینه‌سازی براساس یک تابع خطای مقعر و یک عبارت رگولاریزیشن محدب ارائه می‌شود، که به ویژه زمانی که تعداد برچسب‌های موجود کم هستند کاملا مناسب است واین نوع مسئله با این خطا، بهترین گزینه برای مسائل دسته‌بندی است.

[1] نوعی از یادگیری است که براساس داده‌های دسته‌بندی شده (برچسب خورده)، برچسب داده‌های دسته‌بندی نشده یا همان دسته‌ای که به آن تعلق دارند، را مشخص می‌کنند.

آزمایش درستی الگوریتم

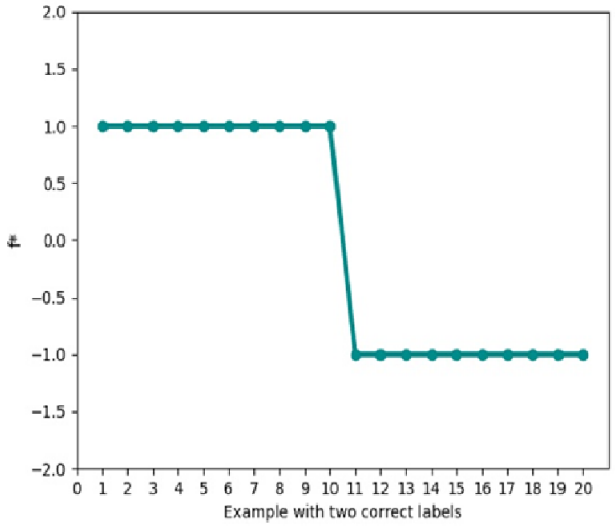
به منظور آزمایش درستی الگوریتم پیاده‌سازی شده مجموعه داده moon که در مقاله [۱] به آن اشاره شده است تولید گردید. این مجموعه داده به صورت یک گراف وزن‌دار است که دارای دو اجتماع با پیوند قوی در آن می‌باشد و این دو اجتماع از طریق یک لینک با پیوند ضعیف (۱۰ برابر ضعیف تر از پیوند درون اجتماع ها) بهم متصل شده‌اند.

۲-۱ آزمایش اول

در آزمایش اول یک گره از کلاس ۱ و یک گره از کلاس -۱ دارای برچسب هستند (گره‌های شماره ۹ و ۱۲) و سایر گره‌ها برچسب ندارند. این مجموعه داده را به عنوان ورودی به الگوریتم می‌دهیم و الگوریتم RobustGC با $\eta = ۰,۵$ باید برچسب گره‌هایی که برچسب ندارند را پیش‌بینی کند. نتایج پیش بینی به صورت زیر است:

```
[ 1.  1.  1.  1.  1.  1.  1.  1.  1.  1. -1. -1. -1. -1. -1. -1. -1. -1.
 -1. -1.]
```

همان‌طور که مشاهده می‌شود الگوریتم توانسته برچسب همه گره‌ها را به درستی پیش‌بینی کند. نمودار برچسب‌های پیش‌بینی شده به ازای هر گره در شکل ۲-۱ زیر آورده شده است.



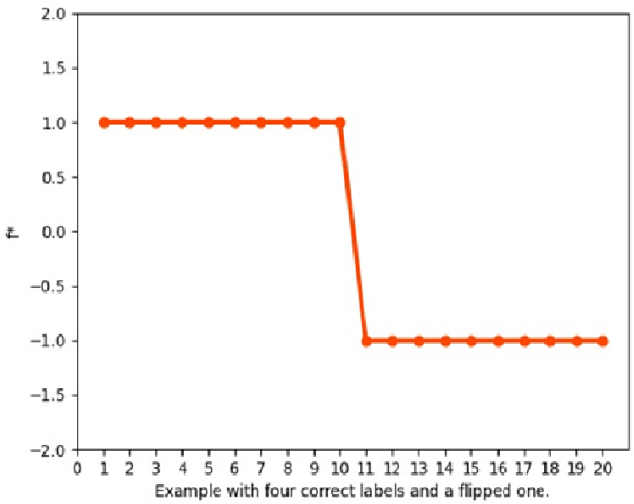
شکل ۲-۱ نمودار برچسب‌های پیش‌بینی شده در آزمایش اول بر روی مجموعه داده moon

۲-۲ آزمایش دوم

در آزمایش دوم یک گره از کلاس ۱ دارای برچسب(گره شماره ۹) است و ۴ گره از کلاس -۱ دارای برچسب هستند که یکی از داده‌های کلاس -۱ به غلط با برچسب ۱ مشخص می‌شود. (گره‌های شماره ۱۲ و ۱۴ و ۱۵ دارای برچسب -۱ و گره شماره ۱۳ دارای برچسب ۱ هستند) نتایج پیش‌بینی به صورت زیر است:

```
[ 1.  1.  1.  1.  1.  1.  1.  1.  1.  1. -1. -1. -1. -1. -1. -1. -1. -1.
 -1. -1.]
```

مجددا الگوریتم توانسته پیش‌بینی برچسب‌ها را کاملا درست انجام بدهد. نمودار برچسب‌های پیش-بینی شده به ازای هر گره در شکل زیر آورده شده است:



شکل ۲-۲ نمودار برچسب های پیش بینی شده در آزمایش دوم بر روی مجموعه داده moon

مجموعه‌داده‌های مورد استفاده برای ارزیابی

- **Digits٤٩-w:** این دیتاست مشابه توضیحی که در مقاله [۱] داده شده است ساخته شد. بدین صورت که از مجموعه داده ارقام USPS تعداد ۱۲۵ عدد رقم ۴ و ۱۲۵ عدد رقم ۹ جدا شد. سپس با استفاده از کرنل گوسی ماتریس وزن برای این مجموعه داده ساخته شد. لیبل‌های داده‌ها نیز به این صورت مشخص گردید که به ازای مشاهده رقم ۴ برچسب -۱ و به ازای مشاهده رقم ۹ برچسب ۱ اختصاص داده شد.

- **Karate**: این مجموعه داده گرافی شامل ۳۴ نود و دو اجتماع گرافی در داخل آن می‌باشد. ماتریس وزن برای این مجموعه داده طبق ارتباط میان نودها (یعنی یال‌ها) تعیین شد و چون وزن هر یال ۱ است در واقع برابر با ماتریس مجاورت گراف مربوط به این مجموعه داده می‌شود. برچسب-های گره‌ها از طریق جستجو در اینترنت یافت شد و به ورودی الگوریتم داده شده است.

- **Dolphins**: این مجموعه داده شامل ۶۲ گره است و ماتریس وزن از طریق ارتباط بین گره‌های آن، با وزن ۱ به ازای هر یال ایجاد می‌شود. شامل دو برچسب ۱ و ۲ است که با ۱- و ۱ جایگزین شد.

- **Polbooks**: آخرین ورژن این دیتاست شامل ۱۰۵ گره و ۳ کلاس با برچسب‌های ۰و۱و۲ است. تعداد گره‌های با برچسب ۰، ۱۳ گره است. با حذف این تعداد از ماتریس وزن و برچسب گره‌ها به تعداد مد نظر مقاله [۱] میرسیم. ماتریس وزن نیز از طریق ارتباطاتی که در فایل دیتاست موجود بود با وزن ۱ به ازای هر یال ساخته شد. جزئیات پیاده‌سازی در بخش polbooks از فایل ضمیمه‌شده موجود است.

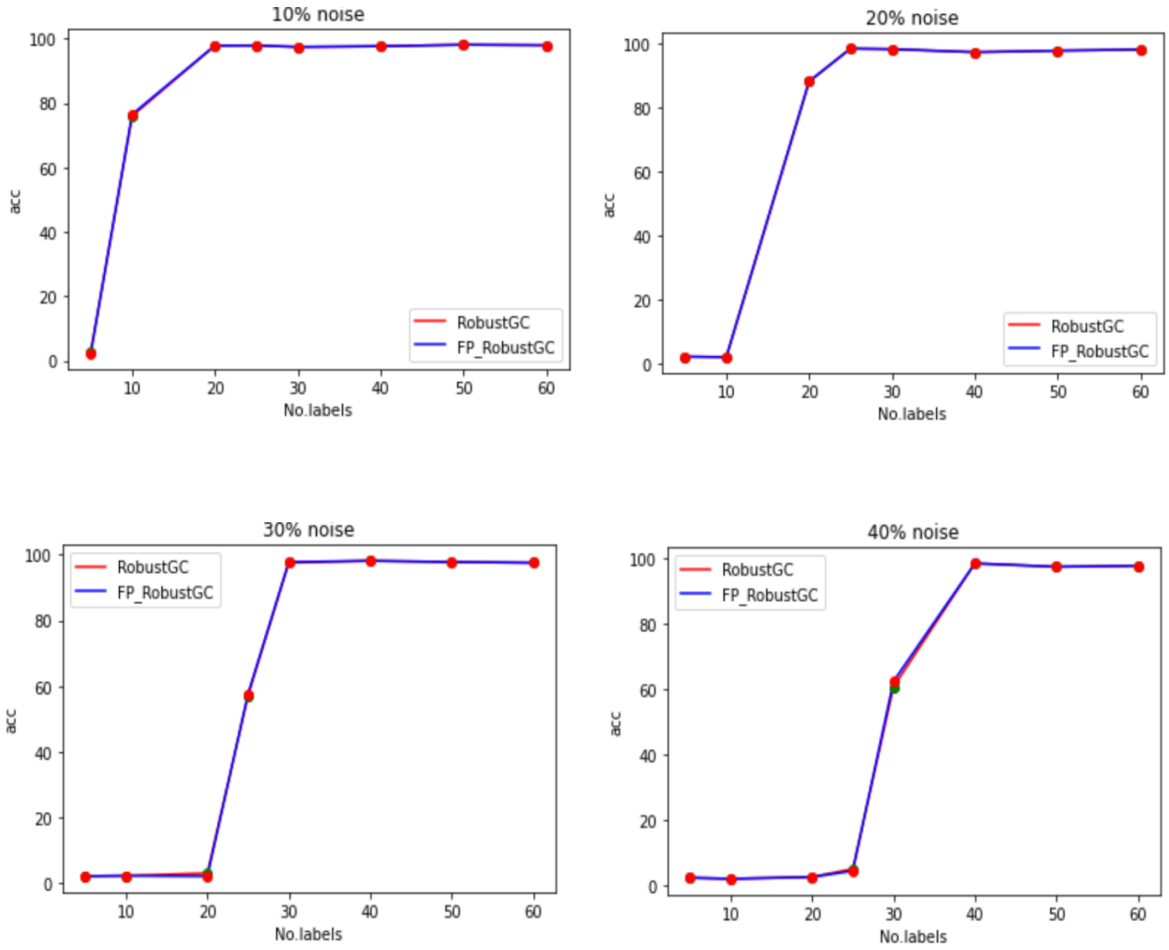
آزمایش‌ها و نتایج

۴–۱ مقاومت الگوریتم در برابر برچسب‌های دارای نویز

در این بخش مقاومت دسته‌بندی الگوریتم‌های RobustGC و FP_RobustGC نسبت به برچسب‌های نویزی سنجیده می‌شود. عملکرد هر دو مدل برای تعداد مختلفی از داده‌های برچسب‌دار و سطوح مختلفی از نویز مقایسه می‌شود. به این صورت که درصدهای مختلفی برای نویز و گره‌های بدون برچسب در نظر گرفته شده و برای پیش‌بینی برچسب گره‌ها، از هر دو مدل استفاده می‌کنیم. هر پیکربندی ۵۰ بار تکرار شده و میانگین دقت برای این تکرارها محاسبه می‌شود. این مراحل را برای هر دو الگوریتم انجام می‌دهیم. مقدار η را برای الگوریتم RobustGC، ۰٫۵ در نظر گرفتیم که مقداری بین ۰ و ۱ است.

۴–۱–۱ نتایج

در صورت استفاده از دیتاست معرفی شده در مقاله [۱] الگوریتم RobustGC، مستقل از سطح نویز اعمال‌شده به برچسب‌ها، توانسته تقریبا همه‌ی گره‌ها را به طور کامل دسته‌بندی کند. الگوریتم FP_RobustGC نیز به خوبی RobustGC است و فقط زمانی که تعداد نویزها بالا و تعداد گره‌های برچسب دار کم باشد، کمی ضعیف تر عمل کرده است. در این جا نیز هر دو الگوریتم به خوبی هم عمل کرده اند و در زمانی که تعداد داده‌های برچسبدار کم است، در سطوح مختلف نویز، عملکردشان به نحوی است که تقریبا تمام نقاط به خوبی دسته‌بندی شده‌اند. نتایج حاصله در شکل ۴-۱ قابل مشاهده است.



شکل ۴-۱ نتایج حاصل از سنجش مقاومت الگوریتم در برابر برچسب های نویزی در دسته‌بندی گره‌ها

۴-۲ سنجش دقت دسته‌بندی

در این بخش برچسب گره‌های بدون برچسب توسط دو الگوریتم RobustGC و FP_RobustGC پیش‌بینی می‌شود. تفاوت این دو الگوریتم در مقدار η است. این مقدار برای FP_RobustGC ۰,۹ و برای RobustGC مقداری بین ۰ و ۱ است که ما ۰,۵ فرض کردیم. برای هر درصدی از داده‌های بدون برچسب و هر مجموعه داده آزمایش ۲۰ بار تکرارشد و میانگین دقت بدست آمد.

۴-۲-۱ نتایج

میانگین دقت‌های بدست آمده به ازای درصدهای مختلف از داده‌های بدون برچسب در جدول ۴-۱ آمده است.

	labels	RobustGC	PF-RobustGC
polbooks	٪۱	–	–
	٪۲	۹۷,۷۷	۹۷,۸۳۳۳
	٪۵	۹۷,۸۷	۹۷,۹۸۸۵
	٪۱۰	۹۸,۰۴۸۷	۹۷,۷۴۳۹
	٪۲۰	۹۸,۱۵۰۶	۹۷,۹۴۵۲
	٪۵۰	۹۸,۱۵۲۱	۹۷,۷۱۷۳
digits۴۹-w	٪۱	۸۷,۴۴۹۳	۸۹,۶۷۶۱
	٪۲	۹۰,۶۷۳۴	۹۲,۳۲۶۵
	٪۵	۹۳,۲۰۶۷	۹۱,۹۸۳۱
	٪۱۰	۹۳,۸۴۴۴	۹۳,۳۷۷۷
	٪۲۰	۹۳,۸	۹۳,۴۲۵
	٪۵۰	۹۴,۹۲	۹۴,۱۱۹۹
karate	-	-	–
	-	-	–
	٪۵	۹۷,۹۶۸۷	۹۷,۶۵۶۲
	٪۱۰	۹۸,۵۰۰۰	۹۷,۶۶۶۶
	٪۲۰	۹۹,۰۷۴۰	۹۷,۷۷۷۷
	٪۵۰	۹۸,۵۲۹۴	۹۶,۷۶۴۷
dolphins	٪۱	۹۶,۷۵۰۰	۹۶,۵۸۳۳
	٪۲	۹۶,۶۶۶۶	۹۶,۶۶۶۶
	٪۵	۹۶,۹۸۲۷	۹۶,۷۲۴۱
	٪۱۰	۹۶,۹۹۹۹	۹۶,۹۰۹۰
	٪۲۰	۹۶,۲۲۴۴	۹۷,۴۴۸۹
	٪۵۰	۹۷,۰۹۶۷۷	۹۶,۴۵۱۶

جدول ۴-۱ میانگین دقت‌های بدست آمده از اعمال ۴ مجموعه داده معرفی شده در الگوریتم‌های RobustGC ، FP_RobustGC

۴–۳ تعمیم به حالت خارج از نمونه[۱]

در این بخش حالت out of sample برای الگوریتم پیاده‌سازی و آزمایش می‌شود. برای به دست آوردن برچسب داده جدیدی که وارد می‌شود، $\text{sign}(f_x)$ را محاسبه می‌کنیم که رابطه f_x طبق مقاله [۱] به این صورت است:

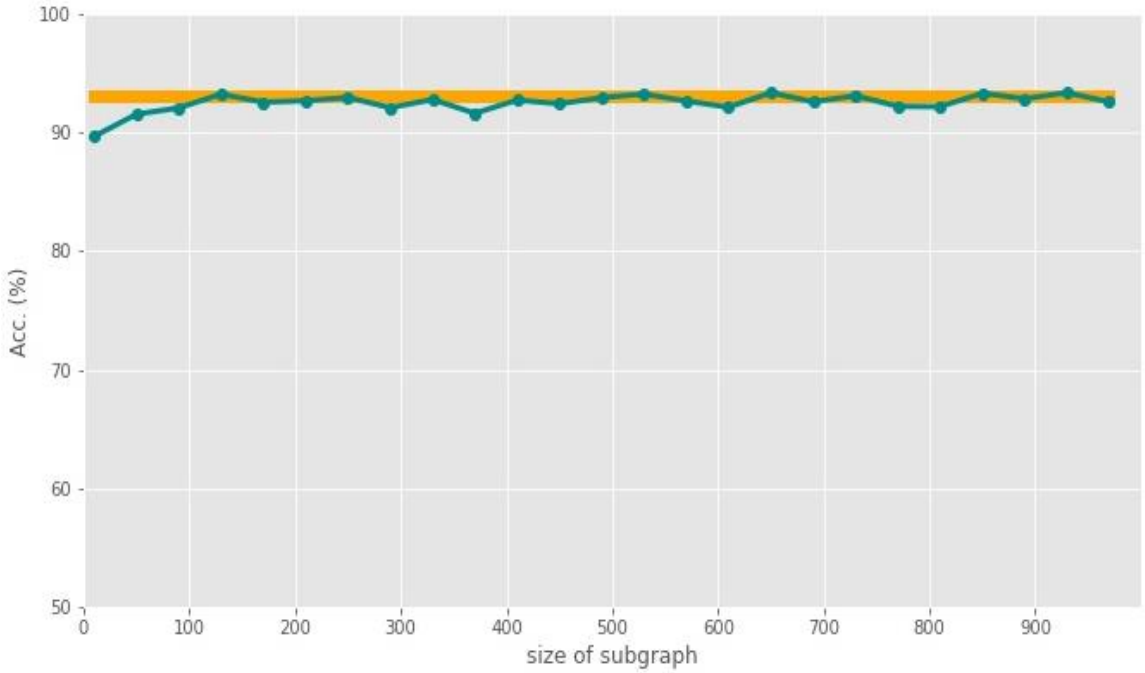
$$f_x = \frac{1}{1-\gamma-k(x,x)d_x^{-1}} \sum_{i=1}^{N} S_{xi} f_i^{\star},$$

آزمایش این بخش بر روی مجموعه داده digits۴۹-w انجام گرفت. تعداد ۱۰۰۰ داده جدا گردید که ۵۰۰ تای آن رقم ۴ و ۵۰۰ تای آن رقم ۹ است. در همه تکرارهای آزمایش، از این مجموعه داده ۱۰۰۰ تایی ۱۰ داده را به صورت برچسب‌دار وارد مجموعه داده آموزش می‌کنیم. اندازه مجموعه داده آموزش متغیر است و از تعداد ۲۰ (۱۰ داده برچسب‌دار و ۱۰ داده بدون برچسب) تا ۱۰۰۰ (۱۰ داده برچسب‌دار و ۹۹۰ داده بدون برچسب) تغییر می‌کند. ماتریس وزن با استفاده از کرنل گوسی برای مجموعه داده آموزشی ساخته می‌شود. سایر داده‌هایی که در مجموعه آموزشی قرار ندارند برچسبشان با استفاده از الگوریتم out of sample پیش‌بینی می‌شود. به دلیل زمان‌بر بودن اجرای این بخش، طول گام برای افزایش تعداد داده‌های آموزش، ۴۰ در نظر گرفته شده است. هم‌چنین دقت در هر مرحله برای ۹۹۰ داده ای که در هر گام برچسب ندارند اندازه‌گیری شده است.

۴–۳–۱ نتایج

نمودار دقت حاصل از اجرای الگوریتم در شکل ۴-۲ آورده شده است. خط نارنجی نشان دهنده دقت به دست آمده به ازای استفاده از ۱۰ داده برچسب‌دار و ۹۹۰ داده بدون برچسب به عنوان داده آموزشی است و هیچ داده ای به عنوان out of sample در این حالت به الگوریتم داده نشده است. همانطور که مشاهده می‌شود دقت های به دست آمده در حالت out of sample بسیار نزدیک به خط نارنجی هستند که نشان می‌دهد الگوریتم در این حالت هم به خوبی عمل کرده است.

[۱] Out-of-Sample Extension



شکل ۴-۲ دقت الگوریتم در حالت out-of-sample بر روی اندازه های مختلف مجموعه داده آموزشی

منابع و مراجع

[۱] M. F. J. A. S. Carlos M. Alaíz, "Robust Classification of Graph-Based Data," *Data Min Knowl Disc ۳۳,* p. ۲۳۰–۲۵۱, ۲۰۱۹.