

مقدمه به تشخیص خطا با استفاده از CRC

تهیه و تنظیم : محمدمین خدادادی - ۴۰۰۱۲۳۴۰۱۱۹۱۶۲ - دکتر مهدی اسلامی - درس شبکه های مخابراتی

مقدمه و بررسی اجمالی: پروژه حاضر با هدف بررسی و تحلیل عملکرد الگوریتم Cyclic Redundancy Check (CRC) در تشخیص خطاها طراحی شده است. یکی از روش های پرکاربرد در سیستم های ارتباطی دیجیتال برای بررسی صحت داده ها است. این الگوریتم بر پایه محاسبات ریاضیاتی مبتنی بر چندجمله ای ها (Polynomials) عمل می کند و قادر است انواع مختلفی از خطاها را شناسایی کند. انواع رایج CRC شامل CRC-8 ، CRC-16 ، CRC-32 و CRC-64 هستند که هر کدام بر اساس درجه چندجمله ای generator خود، ظرفیت تشخیص خطا متفاوتی دارند.

عملکرد CRC: CRC با افزودن یک کد اضافی (checksum) به داده های اصلی کار می کند. این کد بر اساس تقسیم مدول-۲ داده ها بر یک چندجمله ای generator محاسبه می شود. هنگام دریافت داده، گیرنده همان عملیات را تکرار می کند و اگر نتیجه صفر نباشد، خطایی رخ داده است. این روش به ویژه در شبکه های کامپیوتری، ذخیره سازی داده ها و ارتباطات بی سیم کاربرد دارد. برای مثال، در پروتکل های شبکه مانند Ethernet از CRC-32 استفاده می شود تا اطمینان حاصل شود که داده ها بدون تغییر منتقل شده اند.

انواع خطاها قابل تشخیص توسط CRC: CRC قادر است خطاهای مختلفی را با احتمال بالا تشخیص دهد، اما تضمین ۱۰۰٪ ندارد. برخی از خطاهای رایج عبارتند از:

- **Random Bit Flip:** تغییر تصادفی یک یا چند بیت در داده ها، که CRC می تواند آن ها را با احتمال بالا شناسایی کند.
- **Burst Error:** خطاهای پیوسته در یک بلوک از بیت ها، مانند آنچه در کانال های نویزی رخ می دهد CRC-32. و بالاتر معمولاً در تشخیص این نوع خطاها قوی تر هستند.
- **خطاها در کانال های Symmetric Binary:** مدل هایی که احتمال خطا در هر بیت برابر است.

در این پروژه، عملکرد CRC روی مجموعه داده های متنوع ارزیابی می شود تا ببینیم چگونه این الگوریتم در سناریوهای واقعی عمل می کند. نتایج نشان می دهد که CRC می تواند بیش از ۹۹٪ خطاها را در شرایط ایده آل تشخیص دهد، اما در موارد پیچیده تر مانند داده های بزرگ یا نویز بالا، ممکن است collision (تشخیص نادرست) رخ دهد. برای بهبود عملکرد، از CRC های طولانی تر مانند CRC-64 استفاده می شود که فضای کد بزرگتری فراهم می کند و احتمال collision را کاهش می دهد. همچنین، در کاربردهای عملی مانند ذخیره سازی داده ها، CRC با ترکیب با روش های دیگر مانند ECC (Error Correcting Codes) استفاده می شود تا نه تنها خطا را تشخیص دهد، بلکه آن را اصلاح کند.

مزایا و محدودیت ها: مزایای CRC شامل سادگی پیاده سازی، سرعت بالا و مصرف کم منابع است. با این حال، محدودیت هایی مانند عدم توانایی در اصلاح خطا (فقط تشخیص) و حساسیت به نوع generator polynomial وجود دارد. در این پروژه، تمرکز بر ارزیابی تجربی این محدودیت ها است.

۱-۲ انواع رایج CRC

در این پروژه از انواع استاندارد CRC استفاده می شود که هر کدام برای کاربردهای خاصی طراحی شده اند. لیست آن ها عبارت است از:

۱. **CRC-8-ATM:** برای ارتباطات ATM و تشخیص خطا در بسته های کوچک.

۲. **CRC-8-Dallas**: مورد استفاده در سنسورهای دالاس برای داده‌های کوچک مانند دما.

۳. **CRC-16-CCITT**: استاندارد در پروتکل‌های مخابراتی مانند X.25 و Bluetooth.

۴. **CRC-16-MODBUS**: برای ارتباطات صنعتی در پروتکل MODBUS.

۵. **CRC-32-IEEE**: پرکاربرد در Ethernet و ZIP files برای داده‌های بزرگتر.

۶. **CRC-32C**: نسخه بهبودیافته برای ذخیره‌سازی مانند در iSCSI.

۷. **CRC-64-ECMA**: برای داده‌های بسیار بزرگ مانند در فایل‌سیستم‌ها و پایگاه‌داده‌ها.

هر کدام از این انواع بر اساس درجه چندجمله‌ای، طول checksum متفاوتی تولید می‌کنند و در برابر انواع خطاها مقاومت متفاوتی نشان می‌دهند.

۱-۳ روش‌شناسی پروژه

در این پروژه، از ۵ مجموعه‌داده استاندارد از کتابخانه Scikit-Learn برای ارزیابی عملکرد CRC استفاده می‌شود. این داده‌ها با انواع مختلفی از ویژگی‌ها (features) انتخاب شده‌اند تا سناریوهای متنوعی را پوشش دهند. برای هر مجموعه‌داده، خطاهای مصنوعی (مانند bit flip یا burst error) اعمال می‌شود و سپس CRC برای تشخیص آن‌ها تست می‌گردد. از کتابخانه‌های Python مانند NumPy و Pandas برای پردازش داده‌ها و محاسبه CRC استفاده می‌شود.

الف) مجموعه‌داده Iris: این داده‌ها مربوط به طبقه‌بندی گونه‌های گل Iris بر اساس ویژگی‌های فیزیکی است.

- تعداد نمونه‌ها: ۱۵۰

- تعداد ویژگی‌ها: ۴ (طول و عرض کاسبرگ و گلبرگ)

- طبقه‌بندی: ۳ کلاس (Setosa, Versicolor, Virginica)

- نوع داده float: روش ارزیابی: داده‌ها به صورت ساده (بدون پیچیدگی بالا) هستند CRC. برای بررسی تأثیر خطاها روی ویژگی‌های float استفاده می‌شود و نتایج نشان‌دهنده نرخ تشخیص بالا در داده‌های کوچک است.

ب) مجموعه‌داده Digits: این داده‌ها تصاویر دیجیتالی دست‌نویس اعداد (از ۰ تا ۹) را شامل می‌شود.

- تعداد نمونه‌ها: ۱۷۹۷

- تعداد ویژگی‌ها: ۶۴ (هر تصویر ۸×۸ پیکسل)

- طبقه‌بندی: ۱۰ کلاس

- نوع داده integer: روش ارزیابی: تمرکز روی تصاویر و اعمال خطاهای ابعادی CRC. برای شبیه‌سازی نویز در تصاویر استفاده می‌شود و نتایج نشان‌دهنده بهبود تشخیص با CRCهای طولانی‌تر است.

ج) مجموعه‌داده Wine: این داده‌ها مربوط به تحلیل شیمیایی شراب‌های مختلف است.

- تعداد نمونه‌ها: ۱۷۸

- تعداد ویژگی‌ها: ۱۳ (مانند الکل، اسیدیته و غیره)

- نوع داده float :
- طبقه‌بندی: ۳ کلاس روش ارزیابی: ویژگی‌های float پیچیده‌تر هستند CRC. برای بررسی تأثیر نویز روی مقادیر پیوسته استفاده می‌شود و نتایج بر بهبود دقت با افزایش طول CRC تأکید دارد.
- د) مجموعه داده: **Breast Cancer** این داده‌ها ویژگی‌های سلولی تومورهای سینه را برای تشخیص بدخیم/خوش‌خیم شامل می‌شود.

- تعداد نمونه‌ها: ۵۶۹
- تعداد ویژگی‌ها: ۳۰
- نوع داده float :
- طبقه‌بندی: ۲ کلاس (Malignant / Benign) روش ارزیابی: داده‌های پزشکی حساس CRC. برای بررسی نرخ تشخیص خطا در ویژگی‌های ابعادی بالا استفاده می‌شود و نتایج نشان‌دهنده عملکرد خوب در داده‌های واقعی است.
- ه) مجموعه داده: **Synthetic** این داده‌ها با استفاده از تابع `make_classification` از Scikit-Learn ساخته شده‌اند.

- تعداد نمونه‌ها: ۳۰۰
- تعداد ویژگی‌ها: ۱۲
- نوع داده float :
- طبقه‌بندی: ۲ کلاس روش ارزیابی: داده‌های مصنوعی برای کنترل کامل خطاها CRC. برای تست سناریوهای تصادفی استفاده می‌شود و نتایج بر تأثیر پیچیدگی داده‌ها تأکید دارد.

روش تست کلی: برای هر مجموعه داده، انواع CRC (با طول‌های متفاوت) اعمال می‌شود. پیچیدگی داده‌ها (مانند float ساده در Iris ، integer در Digits ، float پیچیده در Wine ، ابعادی بالا در Breast Cancer و مصنوعی در Synthetic) ارزیابی می‌گردد. خطاهای random و burst اعمال شده و نرخ تشخیص محاسبه می‌شود.

۴-۱ نتایج و بحث

در این پروژه، عملکرد CRC-8 ، CRC-16 ، CRC-32 و CRC-64 روی مجموعه داده‌ها ارزیابی شد. نتایج نشان می‌دهد که CRC های طولانی‌تر (مانند CRC-32 و CRC-64) در تشخیص خطاها مؤثرتر هستند، به ویژه در داده‌های بزرگ و پیچیده. نرخ collision در CRC-8 بالاتر است، اما در CRC-64 نزدیک به صفر می‌رسد.

تحلیل عملکرد:

- در برابر random bit flip ، همه CRC ها خوب عمل می‌کنند، اما CRC-64 ۱۰۰٪ تشخیص می‌دهد.
 - در burst errors ، CRC های کوتاه‌تر ضعیف‌تر هستند.
 - در کانال‌های نویزی، ترکیب CRC با روش‌های دیگر پیشنهاد می‌شود.
- نرخ collision کاهش می‌یابد با افزایش طول، و در داده‌های واقعی مانند Breast Cancer ، CRC-32 کافی است. نتایج نشان‌دهنده این است که انتخاب CRC بر اساس اندازه داده و نوع کاربرد حیاتی است.

نتیجه‌گیری CRC: ابزاری قدرتمند برای تشخیص خطا است، اما باید با توجه به محدودیت‌ها استفاده شود. پیشنهاد می‌شود در کاربردهای حساس، از CRC-64 استفاده گردد.

ذخیره نتایج: تمام نتایج در فایل Excel به نام `crc_results_full - v1.xlsx` ذخیره شده است.