**Advanced Intelligent Information Systems –Data Mining**

Is it Possible to classify the popularity of a project using metrics attained from GitHub?

-Jordan McDonald-

**1. Application Problem**

GitHub is a web based GIT repository service that enables version control and source management for teams (distributed or local) to collaborate in software development. GitHub is a platform which thrives upon open source development as well as supporting business applications, at the time of writing GitHub reports having 14 million users and 35 million repositories making it the largest host of source code in the world. This represents a period of rapid growth considering in 2010, announced on the official GitHub blog it was revealed that one million repositories were hosted on GitHub. In the context of data mining this is an application context that is ripe for investigation and has inspired this study.

The problem this report will tackle will be – is it possible to classify the popularity of a repository using only projects that can from obtained from a) the GitHub website b) the GitHub API. GitHub presents a metric called ‘stargazers’ which represents a user who has shown an interest in the project, therefore the more popular a repository is, the higher the amount of stargazers present so this will act as the class attribute which requires a decision. The variance of popularity needs to be quantified in order for this study to succeed, to do this differing sets of stargazers have been devised which provide a multi class range of potential classifications, see below for the ranges.

* <500
* 500-2000
* 2000-5000
* 5000-10000
* >10000

To achieve this goal of classification consideration needs to be given to the types of metrics which can be extracted, from either the API or the website itself, refer to figure one for the metrics that will be utilized and what each mean in the context of the GitHub platform (each metric represents an attribute in the dataset). To ensure the languages selected are no arbitrary, the most popular on the platform have been chosen [1], from this each will be split into subsets according to the various stargazers class attribute ranges.

|  |  |
| --- | --- |
| **Metric** | **Description** |
| Language | This will be a multi-language study and this will represent what programming language is in the instance. |
| Commit | A snapshot of the repository |
| Branches | A separate version of the repository which enables experimentation from outside the master branch. |
| Watchers | Users who subscribe to a repository and receive updates. |
| Forks | Similar to a branch by allowing users to clone the repository and work on it independently. |
| Issues | Users can raise issues when bugs are discovered/suggest areas for improvement. |
| Releases | Represents a version of the system that has been released to the public. |
| LOC | Total lines of code in the repository |
| Pull Requests | A request to merge a change to the master branch |
| Read me? | A file explaining the repository, in this case, does one exist? |
| <100 contributors? | Used to check the size of the development team and see how that impacts popularity. |
| Commits in the last month? | Used to assist in determining the activity rate of the contributors. |
| Stars | Allows a user to register interest in a project |

Figure 1 – metrics used in this study as attributes.

**A) Dataset Generation Strategy**

Now it would be prudent how the data will be collected, GitHub provides an advanced search engine (https://github.com/search/advanced) which allows a user to discover repositories that meet a series of parameters. This was utilized to filter the selection by language and then by then the potential class attribute values, see below for an example search query.

* language:C stars:500..2000 (finds the projects which have 500-200 stars for the C language).

This would result in a list of options that could be selected as part of the dataset, to ensure that no bias was applied the first results were chosen indiscriminately. From this the repository could be visited which presents most of the data required in a visual manner, however in some cases API access will be required. To enable this process the ‘Darwin’ workbench has been leveraged, I developed this as part of my dissertation and decided it would be wise to apply it where required to this scenario, see figure three to view the architecture of this system. The project selection strategy should also be documented, see figure two for the general approach taken.

Identify the top ten languages

Filter the repositories for each language using the search engine

If 12 selected move onto the next language

Select 2-3 repositories for each class attribute range

GitHub API

Webpage

URL(s)

JSON data

Java Servlet

JSON extractor module (JS)

Raw data

MongoDB

DB Query

R Environment

Figure 3 – Darwin system architecture

It should also be considered that this report will use real world data that will be original and extracted by myself, this leads a significant amount of time spent forming the dataset which should be considered when discussing the significance and technical difficulty of the report. This will lead to a reduced scope in the number of instances that are use (120 total) but I feel the originality of the study (which to my knowledge has not been tackled using the two classification techniques chosen) outweighs that restriction.

**B) Description of the dataset**

The dataset will be formatted as an attribute relation file format (arff) and as mentioned prior has been generated using real world repository data. The dataset has been attached in the appendix for future reference. The bulk of the data will be numerical and typically represents counts for a certain metric, so in the case of commits the value for the instance indicates the total amount for that particular repository. Nominal/categorical data is present for three of the attributes with the potential values of ‘Yes’ or ‘No’ which indicate a binary outcome. The class attribute uses ordinal data be presenting various ordered ranges of values with the first and last outcomes being more general using either ‘<’ or ‘>’ to capture the remaining repositories that do not fit within any of the set ranges.

**2. Report Methodology & Classification Model**

//have a section for machine learning techniques, tool (weka), and the two models

**3. Descriptions of Ten Fold Cross Validation & Evaluation**

**References**

* 1. Language Trends on GitHub, https://github.com/blog/2047-language-trends-on-github, August 19, 2015

**Appendix**

@RELATION 'repositories'

@ATTRIBUTE 'Language' {JavaScript, Java, Ruby, PHP, Python, CSS, C++, C#, C, HTML}

@ATTRIBUTE 'Commits' NUMERIC

@ATTRIBUTE 'Branches' NUMERIC

@ATTRIBUTE 'Watchers' NUMERIC

@ATTRIBUTE 'Forks' NUMERIC

@ATTRIBUTE 'Issues' NUMERIC

@ATTRIBUTE 'Releases' NUMERIC

@ATTRIBUTE 'Readme' {Yes, No}

@ATTRIBUTE '<100Contributors' {Yes, No}

@ATTRIBUTE 'LOC' NUMERIC

@ATTRIBUTE 'CommitInTheLastMonth' {Yes, No}

@ATTRIBUTE 'Pull Requests' NUMERIC

@ATTRIBUTE 'stars' {<500, 500-2000, 2000-5000, 5000-10000, >10000}

@data

C,1298,15,97,268,72,25,Yes,Yes,65591,Yes,7,<500 %https://github.com/edenhill/librdkafka

C,24,1,14,53,4,0,Yes,Yes,2389,No,0,<500 %https://github.com/karthick18/inception

C,81,1,107,132,0,0,Yes,Yes,1046,No,0,<500 %https://github.com/x0r1/jellyfish

C,378,127,397,25,15,5,Yes,Yes,6046,No,5,500-2000 %https://github.com/pedrovgs/DraggablePanel

C,104,2,132,179,52,7,Yes,Yes,9627,No,0,500-2000 %https://github.com/google/ios-webkit-debug-proxy

C,677,22,427,1209,58,8,Yes,Yes,15976,No,5,2000-5000 %https://github.com/square/dagger

C,11304,7,426,2355,387,0,Yes,No,321560,Yes,9,2000-5000 %https://github.com/codecombat/codecombat

C,266,7,600,3400,66,8,Yes,Yes,3927,No,43,5000-10000 %https://github.com/scottjehl/Respond

C,3,3,314,81415,657,0,Yes,Yes,46,No,5000,5000-10000 %https://github.com/octocat/Spoon-Knife

C,431,3,461,1785,22,42,Yes,Yes,4181,No,5,5000-10000 %https://github.com/desandro/masonry

C,4017,2,3621,12679,7,0,Yes,No,5730,Yes,13,>10000 %https://github.com/vhf/free-programming-books

C,3521,35,2355,13228,135,192,Yes,No,51675,Yes,66,>10000 %https://github.com/mbostock/d3

JavaScript,451,3,37,37,17,22,Yes,Yes,8946,No,1,<500 %https://github.com/conveyal/transitive.js

JavaScript,122,2,20,54,12,2,Yes,Yes,3446,No,3,<500 %https://github.com/ded/morpheus

JavaScript,95,3,52,99,6,13,Yes,Yes,1694,No,2,500-2000 %https://github.com/stutrek/scrollMonitor

JavaScript,155,3,114,257,14,0,Yes,Yes,101582,No,1,500-2000 %https://github.com/MicrosoftEdge/static-code-scan

JavaScript,175,9,44,50,15,0,Yes,Yes,15080,No,0,500-2000 %https://github.com/begriffs/css-ratiocinator

JavaScript,2565,9,325,1748,252,13,Yes,No,21811,No,26,2000-5000 %https://github.com/novus/nvd3

JavaScript,845,2,342,1258,328,17,Yes,Yes,41839,No,123,2000-5000 %https://github.com/flot/flot

JavaScript,563,4,291,939,467,15,Yes,Yes,261134,No,30,2000-5000 %https://github.com/vitalets/x-editable

JavaScript,266,7,600,3438,66,8,Yes,Yes,3927,No,43,5000-10000 %https://github.com/scottjehl/Respond

JavaScript,431,3,461,1785,22,42,Yes,Yes,4181,No,5,5000-10000 %https://github.com/desandro/masonry

JavaScript,6545,15,3002,6879,449,44,Yes,No,134373,Yes,30,>10000 %https://github.com/facebook/react

JavaScript,6078,4,3251,10436,57,138,Yes,No,62959,Yes,30,>10000 %https://github.com/jquery/jquery

Java,64,1,13,28,14,2,Yes,Yes,3726,No,1,<500 %https://github.com/nzakas/cssembed

Java,51,3,17,71,1,0,Yes,Yes,2442,No,1,<500 %https://github.com/JorgeCastilloPrz/ExpandablePanel

Java,185,7,524,1098,31,21,Yes,Yes,234441,Yes,3,500-2000 %https://github.com/alibaba/jstorm

Java,509,2,135,485,45,23,Yes,Yes,30199,Yes,1,500-2000 %https://github.com/code-troopers/android-betterpickers

Java,677,22,427,1209,58,8,Yes,Yes,15979,No,5,2000-5000 %https://github.com/square/dagger

Java,31202,639,354,534,0,5212,Yes,No,1498133,Yes,20,2000-5000 %https://github.com/JetBrains/kotlin

Java,73,5,270,847,3,0,Yes,Yes,3269,Yes,3,2000-5000 %https://github.com/lgvalle/Material-Animations

Java,891,58,825,2629,138,20,Yes,Yes,15431,Yes,28,5000-10000 %https://github.com/square/picasso

Java,203,1,1105,1789,25,1,Yes,Yes,5352,Yes,3,5000-10000 %https://github.com/futurice/android-best-practices

Java,405,5,841,2579,62,5,Yes,Yes,9834,Yes,3,5000-10000 %https://github.com/greenrobot/EventBus

Java,21475,53,1622,5417,1008,148,Yes,No,1121692,Yes,3,>10000 https://github.com/elastic/elasticsearch

Java,4670,4,1125,2117,101,138,Yes,No,131796,Yes,3,>10000 %https://github.com/ReactiveX/RxJava

Ruby,342,5,34,296,34,22,Yes,No,11279,No,3,<500 %https://github.com/cschiewek/devise\_ldap\_authenticatable

Ruby,161,1,21,107,14,9,Yes,Yes,3453,No,3,<500 %https://github.com/maxjustus/sinatra-authentication

Ruby,1086,7,67,153,12,92,Yes,Yes,6753,No,5,500-2000 %https://github.com/wvanbergen/request-log-analyzer

Ruby,4670,4,1125,2117,101,138,Yes,No,8409,Yes,3,500-2000 %https://github.com/intridea/hashie

Ruby,6587,71,250,1072,33,71,Yes,No,11232,Yes,3,2000-5000 %https://github.com/cucumber/cucumber-ruby

Ruby,246,3,200,286,12,9,Yes,Yes,1810,No,4,2000-5000 %https://github.com/rails-api/rails-api

Ruby,43056,18,1022,2794,123,309,Yes,Yes,13564,Yes,3,5000-10000 %https://github.com/ruby/ruby

Ruby,857,16,414,3119,180,1,Yes,No,3222,No,98,5000-10000 %https://github.com/imathis/octopress

Ruby,49212,2,217,3517,63,114,Yes,No,223442,Yes,15,5000-10000 %https://github.com/caskroom/homebrew-cask

Ruby,57373,22,2214,12487,418,290,Yes,No,394621,Yes,508,>10000 %https://github.com/rails/rails

Ruby,7126,25,1269,5346,138,97,Yes,No,234165,Yes,3,>10000 %https://github.com/jekyll/jekyll

Ruby,18715,15,817,4767,0,159,Yes,No,88721,Yes,22,>10000 %https://github.com/discourse/discourse

PHP,448,2,33,47,2,33,Yes,Yes,?,No,1342,<500 %https://github.com/peteboere/css-crush

PHP,338,3,75,67,4,0,Yes,Yes,563,No,2,<500 %https://github.com/FriendsOfPHP/security-advisories

PHP,1344,3,90,175,120,37,Yes,Yes,5642,No,10,<500 %https://github.com/rocketeers/rocketeer

PHP,25343,5,324,2143,267,133,Yes,No,1018308,No,332,500-2000 %https://github.com/joomla/joomla-cms

PHP,586,5,101,381,84,70,Yes,Yes,7362,Yes,1,500-2000 %https://github.com/Maatwebsite/Laravel-Excel

PHP,40,1,287,150,1,0,Yes,Yes,2789,No,1,500-2000 %https://github.com/phptodayorg/php-must-watch

PHP,1002,11,244,450,128,2,Yes,Yes,44321,No,6,2000-5000 %https://github.com/twostairs/paperwork

PHP,1490,2,278,851,35,32,Yes,No,20410,No,10,2000-5000 %https://github.com/Seldaek/monolog

PHP,601,4,920,1930,4,0,Yes,Yes,8911,Yes,4,5000-10000 %https://github.com/domnikl/DesignPatternsPHP

PHP,1625,5,432,1252,14,6,Yes,No,13,Yes,33,5000-10000 %https://github.com/fzaninotto/Faker

PHP,4804,5,3097,7517,0,63,Yes,No,45,Yes,0,>10000 %https://github.com/laravel/laravel

PHP,25926,11,1068,4788,572,164,Yes,No,123,Yes,155,>10000 %https://github.com/symfony/symfony

Python,229,2,48,97,38,33,Yes,Yes,73054,No,0,<500 %https://github.com/sametmax/0bin

Python,2133,17,45,53,1,1,Yes,Yes,9002,No,0,<500 %https://github.com/guardian/alerta

Python,998,2,168,539,79,1,Yes,Yes,2114,No,23,500-2000 %https://github.com/Lasagne/Lasagne

Python,812,7,179,216,48,32,Yes,Yes,8321,No,48,500-2000 %https://github.com/spinnaker/spinnaker

Python,338,3,18,63,182,25,Yes,Yes,17343,No,0,500-2000 %https://github.com/kennethreitz/clint

Python,448,2,611,1596,163,23,Yes,Yes,20011,No,12,2000-5000 %https://github.com/fxsjy/jieba

Python,2207,3,293,364,53,45,Yes,Yes,57176,No,1,2000-5000 %https://github.com/nicolargo/glances

Python,1970,3,242,373,84,47,Yes,No,16321,No,31,2000-5000 %https://github.com/nate-parrott/Flashlight

Python,9,1,19,29,0,0,Yes,Yes,?,No,1837,<500 %https://github.com/mschwager/dhcpwn

Python,138,1,86,250,19,9,Yes,Yes,8332,No,48,500-2000 %https://github.com/JakeWharton/pidcat

Python,9493,31,287,1573,349,158,Yes,No,22333,Yes,31,2000-5000 %https://github.com/celery/celery

Python,261,7,293,364,19,5,Yes,Yes,10964,No,1,2000-5000 %https://github.com/facebook/chisel

CSS,84,2,40,88,0,2,Yes,Yes,343,No,0,<500 %https://github.com/kogakure/gitweb-theme

CSS,3,1,31,91,0,0,Yes,Yes,1231,No,0,<500 %https://github.com/codrops/PageLoadingEffects

CSS,81,1,236,145,1,0,Yes,Yes,765,No,0,500-2000 %https://github.com/AllThingsSmitty/must-watch-css

CSS,745,3,236,638,11,0,Yes,Yes,896,No,0,500-2000 %https://github.com/1sters/material\_design\_zh

CSS,233,1,60,190,32,10,Yes,Yes,123,No,0,500-2000 %https://github.com/rstacruz/flatdoc

CSS,1178,4,62,209,9,3,Yes,Yes,8641,No,2,500-2000 %https://github.com/mdo/github-buttons

CSS,30,1,80,49,7,0,Yes,Yes,78232,No,0,<500 %https://github.com/m242/maildrop

CSS,306,2,47,55,7,18,Yes,Yes,4832,No,0,<500 %https://github.com/HubSpot/tooltip

CSS,47,2,20,32,4,2,Yes,Yes,22453,No,2,<500 %https://github.com/bjork24/Unison

CSS,178,4,62,209,9,3,Yes,Yes,2890,No,2,500-2000 %https://github.com/mdo/github-buttons

CSS,210,2,48,180,6,31,Yes,Yes,2012,No,1,500-2000 %https://github.com/wavded/humane-js

CSS,210,2,48,180,6,31,Yes,Yes,6652,No,1,500-2000 %https://github.com/poole/poole

C++,144,4,22,78,12,24,Yes,Yes,33818,No,3,<500 %https://github.com/node-inspector/v8-profiler

C++,1540,1,29,82,47,0,Yes,Yes,28901,No,1,<500 %https://github.com/etexteditor/e

C++,209,1,71,86,0,207,Yes,Yes,106090,No,0,<500 %https://github.com/AutoHotkey/AutoHotkey

C++,135,2,156,148,6,2,Yes,Yes,248112,Yes,0,500-2000 %https://github.com/electronicarts/EASTL

C++,128,5,189,548,33,0,Yes,Yes,8921,No,9,500-2000 %https://github.com/codebutler/firesheep

C++,2826,2,213,391,96,0,Yes,Yes,183712,No,9,2000-5000 %https://github.com/paulasmuth/fnordmetric

C++,39074,13,497,3461,221,110,Yes,Yes,99221,Yes,221,2000-5000 %https://github.com/xbmc/xbmc

C++,2052,1,318,444,113,7,Yes,Yes,68998,Yes,19,2000-5000 %https://github.com/SFTtech/openage

C++,3692,7,1332,5709,396,11,Yes,No,144659,Yes,211,5000-10000 %https://github.com/BVLC/caffe

C++,18884,2,1105,7890,942,48,Yes,No,111234,Yes,58,5000-10000 %https://github.com/Itseez/opencv

C++,36373,17,1972,3127,1643,68,Yes,Yes,637281,Yes,15,>10000 %https://github.com/apple/swift

C#,488,2,49,233,41,0,Yes,Yes,831,No,23,<500 %https://github.com/migueldeicaza/MonoTouch.Dialog

C#,8,1,64,131,4,0,Yes,Yes,11022,No,5,<500 %https://github.com/wybory2014/Kalkulator1

C#,3348,15,162,270,312,32,Yes,Yes,146868,Yes,5,500-2000 %https://github.com/jaredpar/VsVim

C#,4029,9,249,885,58,50,Yes,Yes,293102,Yes,4,500-2000 %https://github.com/cefsharp/CefSharp

C#,13404,41,793,1248,2611,35,Yes,Yes,65212,Yes,107,2000-5000 %https://github.com/dotnet/roslyn

C#,106398,219,517,2171,0,316,Yes,Yes,29321,Yes,111,2000-5000 %https://github.com/mono/mono

C#,10041,10,1433,2274,938,3,Yes,Yes,2393113,Yes,39,5000-10000 %https://github.com/dotnet/corefx

C#,5212,3,978,710,0,1,Yes,Yes,647212,Yes,79,5000-10000 %https://github.com/dotnet/coreclr

C#,4520,92,667,1732,397,36,Yes,Yes,85391,No,18,5000-10000 %https://github.com/SignalR/SignalR

C#,175,3,53,90,39,10,Yes,Yes,4492,No,2,<500 %https://github.com/kohsuke/winsw

C#,1747,3,75,65,66,6,Yes,Yes,17789,Yes,0,<500%https://github.com/Pash-Project/Pash

C#,1612,3,131,227,238,26,Yes,Yes,5226,Yes,11,>10000%https://github.com/chocolatey/choco

HTML,66,1,38,80,20,4,Yes,Yes,7890,No,2,<500 %https://github.com/scotch-io/scotch-panels

HTML,15,1,44,108,0,0,Yes,Yes,14016,No,3,<500 %https://github.com/Aaaaaashu/Front-End-Style-Guide

HTML,586,2,57,283,35,0,Yes,Yes,88221,Yes,2,<500 %https://github.com/coursera-dl/edx-dl

HTML,373,13,121,641,120,19,Yes,Yes,9021,No,43,500-2000 %https://github.com/vitch/jScrollPane

HTML,248,3,94,94,0,15,Yes,Yes,67821,Yes,0,500-2000 %https://github.com/caiorss/Functional-Programming

HTML,203,2,75,98,19,13,Yes,Yes,13411,No,2,500-2000 %https://github.com/micha/resty

HTML,160,2,141,432,11,6,Yes,Yes,864,No,4,500-2000 %https://github.com/sofish/typo.css

HTML,482,4,144,443,60,27,Yes,Yes,7212,Yes,21,500-2000 %https://github.com/grangier/python-goose

HTML,26,1,49,78,3,8,Yes,Yes,?,No,2001,500-2000 %https://github.com/1000ch/grd

HTML,611,36,124,192,26,2,Yes,Yes,65120,No,17,<500 %https://github.com/strangeioc/strangeioc

HTML,338,2,28,358,11,9,Yes,Yes,890,Yes,2,<500 %https://github.com/Huxpro/huxpro.github.io

HTML,122,5,26,36,0,6,Yes,Yes,1181,No,0,<500 %https://github.com/bensmithett/style