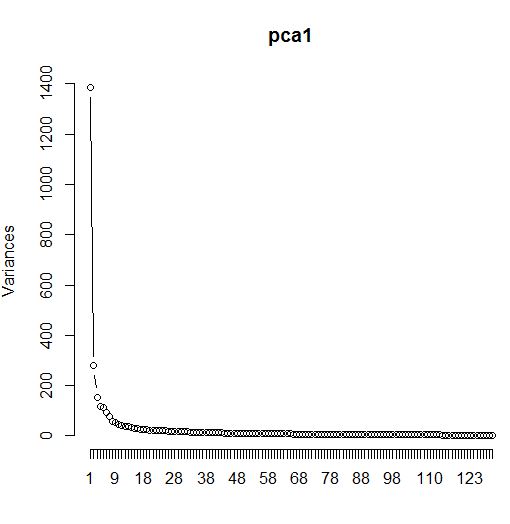
**Feature selection from new malware samples**

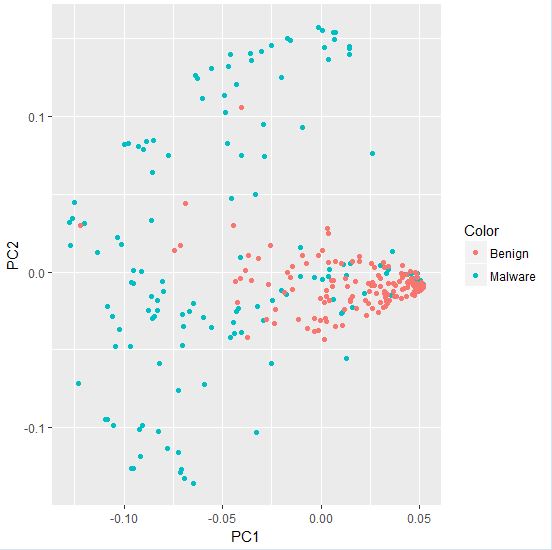
To best of our knowledge, no categorized samples like Genome dataset is available for new malware samples. Therefore, for new malware samples, identifying a malicious service in a malware sample and extracting its API calls for classification like the approach done in experiment 1 is very time consuming and almost impossible.

In the following experiment, we want to do classification of new malware samples and benign samples based on their service behavior whereas it is not determined which one of their services host malicious behavior. In this regard, we used the union of API calls extracted from the services in an application to do classification. For example, if malware sample M1 contains three services, S1, S2 and S3. We extracted the API calls through all these services’ lifecycle and unit them as the API call traces of M1. Note that there may be only either one of the services or more or even none of them contains malicious operation. As mentioned earlier, it is because of repackaged nature of malware samples. We are trying to do kindly blackbox testing of classification based on service behaviors. We carried out making union of API call traces of services for all samples in Androzoo dataset and GooglePlay dataset.

For classification, like the approach presented in figure 1, we extract the itemsets from API call traces using Eclat with specified “support” and “length”. Here, in the following experiment, we extracted the itemsets in Androzoo dataset as well as GooglePlay dataset with the support of 30 percent and the length of 2. As it is expected, all the itemsets extracted form GooglePlay dataset is also extracted in Androzoo dataset. It is because of benign services exist in each malware samples.

We come by 19787 itemsets for support of 30 percent and the length of 2. Using them as classification feature needs high computation time and memory. Therefore, we applied Principle Component Analysis (PCA) [52] for feature reduction purpose. PCA is orthogonal linear transformation of features to a new coordinate system, called components, such that the greatest variance reach by the projection of original feature on each component. For Androozoo itemsets after applying PCA, we got 130 components as new feature space while retaining 95% of cumulative proportion of variance. Figure 1 shows the scree plot, variance retained by each component. We plot the position of applications in the coordinate system made by the first and second principle component having the greatest variance, in Figure 2. It illustrates that benign applications are located in similar area while malware samples distributed in component space. It depict the deviated features in malware samples. In fact, it shows the different behavior exist in services in malware samples.





For classification, we applied 10 fold cross validation on two class classification algorithms, Random Forest, decision Tree and SVM. We kept 30 percent of dataset for classification testing and table 3 depict the result of classification on test data.

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | Accuracy | TPR | FPR |
| Random Forest | 92.55% | 89.28% | 6.89% |
| Decision Tree | 82.33% | 78.57% | 13.78% |
| SVM | 80.17% | 89.28% | 65.51% |

1. Conclusion and Future Work

In this paper, we studied the service components of Android application and showed that malicious operations in malware samples are most likely to be located in service components. We proposed a classification approach by training the classification algorithms by considering the itemset of APIs called through the life cycle of service components in applications. We reached the accuracy of 96% while the FPR is as low as 5%. Whereas our approach does not outperform the detection accuracy of MaMaDroid[24] or APIDroidMiner[29], it needs less computation time since it only do static analysis over service components of applications. All the approaches utilizing machine-learning algorithms need further investigation over samples that did not classify by high probability. Therefore, we believe that our approach needs less effort when we target filtering out the highly safe applications or filtering out the highly probable malicious applications.

The static analysis approach used for API call extraction, in this paper, FlowDroid[49], does not handle reflection calls. Using approaches that consider reflection calls for API call extraction such as our approach proposed in [] will provide accurate results that we will do it as our future work.

> ##############READ DATA

> Trace <-read.transactions("C:/Ava/android/Trace/theZoo2.csv", sep=",") #theZoo2.csv", sep="," , rm.duplicates=TRUE)#APItracesServices.csv", sep=",") #myTrace4.csv", sep=",")# it is malware traces

Warning message:

In asMethod(object) : removing duplicated items in transactions

> TraceGoogleplay <-read.transactions("C:/Ava/android/Trace/GooglePlayTracesC2.csv", sep=",")

Warning message:

In asMethod(object) : removing duplicated items in transactions

>

> itemsets<-eclat(Trace,parameter=list(sup=0.3, minlen=2,maxlen=2))

Eclat

parameter specification:

tidLists support minlen maxlen target ext

FALSE 0.3 2 2 frequent itemsets FALSE

algorithmic control:

sparse sort verbose

7 -2 TRUE

Absolute minimum support count: 51

create itemset ...

set transactions ...[2947 item(s), 171 transaction(s)] done [0.03s].

sorting and recoding items ... [296 item(s)] done [0.00s].

creating bit matrix ... [296 row(s), 171 column(s)] done [0.00s].

writing ... [24594 set(s)] done [0.01s].

Creating S4 object ... done [0.00s].

> itemsetsGooglePlay<- eclat(TraceGoogleplay ,parameter=list(sup=0.3,minlen=2,maxlen=2))

Eclat

> itemsets<-setdiff(itemsets,itemsetsGooglePlay)

>

> ISmatrixZooInGoogleplay <- is.subset(itemsets,TraceGoogleplay)

> tISmatrixZooInGoogleplay<-t(ISmatrixZooInGoogleplay)

>

> ISmatrixTraceInZoo <- is.subset(itemsets,Trace)

> tISmatrixTraceInZoo<-t(ISmatrixTraceInZoo)

> dim(tISmatrixTraceInZoo)

[1] 171 19787

> dim(tISmatrixZooInGoogleplay)

[1] 200 19787

> ISmatrixTraceInZoo <- is.subset(itemsets,Trace)

> tISmatrixTraceInZoo<-t(ISmatrixTraceInZoo)

>

> #add result coulumn

> C = matrix( ,nrow=nrow(tISmatrixTraceInZoo), ncol=1)

> B = matrix( ,nrow=nrow(tISmatrixTraceInZoo), ncol=1)

> colnames(C) <- c("Result")

> colnames(B) <- c("Color")

> for(i in 1:nrow(tISmatrixTraceInZoo))

+ {

+ C[i, ]<-1

+ B[i, ]<-'Malware'

+

+ }

>

> cISmatrixTraceInZoo<-cbind(tISmatrixTraceInZoo, B)

> tISmatrixTraceInZoo<-cbind(tISmatrixTraceInZoo, C)

>

> C = matrix( ,nrow=nrow(tISmatrixZooInGoogleplay), ncol=1)

> B = matrix( ,nrow=nrow(tISmatrixZooInGoogleplay), ncol=1)

>

> colnames(C) <- c("Result")

> colnames(B) <- c("Color")

>

> for(i in 1:nrow(tISmatrixZooInGoogleplay))

+ {

+ C[i, ]<-0

+ B[i, ]<-'Benign'

+ }

> cISmatrixZooInGoogleplay<-cbind(tISmatrixZooInGoogleplay, B)

> tISmatrixZooInGoogleplay<-cbind(tISmatrixZooInGoogleplay, C)

>

> ###############bind two data, malware and Normal

> ZooItemGoogle<-rbind(tISmatrixTraceInZoo,tISmatrixZooInGoogleplay)

> cZooItemGoogle<-rbind(cISmatrixTraceInZoo,cISmatrixZooInGoogleplay)

> my\_data<- subset(ZooItemGoogle, select = -c(Result))

> pca1<-prcomp(my\_data, scores=TRUE, cor=TRUE)

> summary(pca1)

Importance of components:

PC1 PC2 PC3 PC4 PC5 PC6

Standard deviation 37.2187 16.67887 12.41101 10.71770 10.57793 9.63751

Proportion of Variance 0.3713 0.07456 0.04128 0.03079 0.02999 0.02489

Cumulative Proportion 0.3713 0.44583 0.48711 0.51790 0.54789 0.57278

PC7 PC8 PC9 PC10 PC11 PC12 PC13

Standard deviation 8.70862 7.42861 7.35379 6.85444 6.52479 6.16739 5.97722

Proportion of Variance 0.02033 0.01479 0.01449 0.01259 0.01141 0.01019 0.00958

Cumulative Proportion 0.59311 0.60790 0.62239 0.63498 0.64639 0.65659 0.66616

PC14 PC15 PC16 PC17 PC18 PC19 PC20

Standard deviation 5.66972 5.40868 5.38752 5.11265 4.94239 4.87426 4.75360

Proportion of Variance 0.00862 0.00784 0.00778 0.00701 0.00655 0.00637 0.00606

Cumulative Proportion 0.67478 0.68262 0.69040 0.69740 0.70395 0.71032 0.71638

PC21 PC22 PC23 PC24 PC25 PC26 PC27

Standard deviation 4.72173 4.65562 4.56886 4.49861 4.4467 4.36763 4.25361

Proportion of Variance 0.00598 0.00581 0.00559 0.00542 0.0053 0.00511 0.00485

Cumulative Proportion 0.72235 0.72816 0.73375 0.73918 0.7445 0.74959 0.75444

PC28 PC29 PC30 PC31 PC32 PC33 PC34

Standard deviation 4.22688 4.12488 4.07505 4.05795 4.00055 3.87677 3.84025

Proportion of Variance 0.00479 0.00456 0.00445 0.00441 0.00429 0.00403 0.00395

Cumulative Proportion 0.75923 0.76379 0.76824 0.77265 0.77694 0.78097 0.78492

PC35 PC36 PC37 PC38 PC39 PC40 PC41

Standard deviation 3.80636 3.75383 3.68114 3.66924 3.63039 3.5592 3.52736

Proportion of Variance 0.00388 0.00378 0.00363 0.00361 0.00353 0.0034 0.00333

Cumulative Proportion 0.78881 0.79258 0.79622 0.79982 0.80336 0.8067 0.81009

PC42 PC43 PC44 PC45 PC46 PC47 PC48

Standard deviation 3.46659 3.40744 3.35085 3.32442 3.31435 3.2907 3.21859

Proportion of Variance 0.00322 0.00311 0.00301 0.00296 0.00294 0.0029 0.00278

Cumulative Proportion 0.81331 0.81642 0.81943 0.82239 0.82533 0.8282 0.83101

PC49 PC50 PC51 PC52 PC53 PC54 PC55

Standard deviation 3.17079 3.16356 3.14411 3.10307 3.06040 3.04285 3.02339

Proportion of Variance 0.00269 0.00268 0.00265 0.00258 0.00251 0.00248 0.00245

Cumulative Proportion 0.83371 0.83639 0.83904 0.84162 0.84413 0.84661 0.84906

PC56 PC57 PC58 PC59 PC60 PC61 PC62

Standard deviation 3.01693 2.95105 2.93911 2.88464 2.87355 2.84532 2.82035

Proportion of Variance 0.00244 0.00233 0.00232 0.00223 0.00221 0.00217 0.00213

Cumulative Proportion 0.85150 0.85384 0.85615 0.85838 0.86059 0.86276 0.86490

PC63 PC64 PC65 PC66 PC67 PC68 PC69

Standard deviation 2.76489 2.75112 2.74017 2.71545 2.69300 2.68595 2.65802

Proportion of Variance 0.00205 0.00203 0.00201 0.00198 0.00194 0.00193 0.00189

Cumulative Proportion 0.86694 0.86897 0.87099 0.87296 0.87491 0.87684 0.87873

PC70 PC71 PC72 PC73 PC74 PC75 PC76

Standard deviation 2.61349 2.61218 2.56443 2.55456 2.53112 2.51243 2.48727

Proportion of Variance 0.00183 0.00183 0.00176 0.00175 0.00172 0.00169 0.00166

Cumulative Proportion 0.88056 0.88239 0.88415 0.88590 0.88762 0.88931 0.89097

PC77 PC78 PC79 PC80 PC81 PC82 PC83

Standard deviation 2.45813 2.4430 2.43306 2.39181 2.37988 2.37448 2.3618

Proportion of Variance 0.00162 0.0016 0.00159 0.00153 0.00152 0.00151 0.0015

Cumulative Proportion 0.89259 0.8942 0.89578 0.89731 0.89883 0.90034 0.9018

PC84 PC85 PC86 PC87 PC88 PC89 PC90

Standard deviation 2.34021 2.31365 2.29894 2.2861 2.25943 2.24689 2.22059

Proportion of Variance 0.00147 0.00143 0.00142 0.0014 0.00137 0.00135 0.00132

Cumulative Proportion 0.90330 0.90474 0.90615 0.9075 0.90892 0.91027 0.91160

PC91 PC92 PC93 PC94 PC95 PC96 PC97

Standard deviation 2.17993 2.17012 2.16401 2.14777 2.12313 2.10164 2.08500

Proportion of Variance 0.00127 0.00126 0.00126 0.00124 0.00121 0.00118 0.00117

Cumulative Proportion 0.91287 0.91413 0.91539 0.91662 0.91783 0.91902 0.92018

PC98 PC99 PC100 PC101 PC102 PC103 PC104

Standard deviation 2.06246 2.05477 2.04863 2.0266 2.01343 1.97735 1.97275

Proportion of Variance 0.00114 0.00113 0.00112 0.0011 0.00109 0.00105 0.00104

Cumulative Proportion 0.92132 0.92245 0.92358 0.9247 0.92576 0.92681 0.92786

PC105 PC106 PC107 PC108 PC109 PC110 PC111

Standard deviation 1.96391 1.95232 1.94193 1.91526 1.90767 1.89248 1.88829

Proportion of Variance 0.00103 0.00102 0.00101 0.00098 0.00098 0.00096 0.00096

Cumulative Proportion 0.92889 0.92991 0.93092 0.93190 0.93288 0.93384 0.93480

PC112 PC113 PC114 PC115 PC116 PC117 PC118

Standard deviation 1.87265 1.85980 1.83855 1.82579 1.81453 1.80197 1.78843

Proportion of Variance 0.00094 0.00093 0.00091 0.00089 0.00088 0.00087 0.00086

Cumulative Proportion 0.93574 0.93666 0.93757 0.93846 0.93934 0.94021 0.94107

PC119 PC120 PC121 PC122 PC123 PC124 PC125

Standard deviation 1.77635 1.75982 1.74187 1.74047 1.7322 1.71551 1.70404

Proportion of Variance 0.00085 0.00083 0.00081 0.00081 0.0008 0.00079 0.00078

Cumulative Proportion 0.94192 0.94275 0.94356 0.94437 0.9452 0.94597 0.94674

PC126 PC127 PC128 PC129 PC130 PC131 PC132

Standard deviation 1.68867 1.68057 1.67446 1.66632 1.65464 1.65258 1.62941

Proportion of Variance 0.00076 0.00076 0.00075 0.00074 0.00073 0.00073 0.00071

Cumulative Proportion 0.94751 0.94827 0.94902 0.94976 0.95049 0.95123 0.95194

PC133 PC134 PC135 PC136 PC137 PC138 PC139

Standard deviation 1.6184 1.60134 1.59450 1.58981 1.58061 1.56870 1.55990

Proportion of Variance 0.0007 0.00069 0.00068 0.00068 0.00067 0.00066 0.00065

Cumulative Proportion 0.9526 0.95333 0.95401 0.95469 0.95536 0.95602 0.95667

PC140 PC141 PC142 PC143 PC144 PC145 PC146

Standard deviation 1.55352 1.53687 1.52912 1.52651 1.50872 1.4979 1.4970

Proportion of Variance 0.00065 0.00063 0.00063 0.00062 0.00061 0.0006 0.0006

Cumulative Proportion 0.95731 0.95795 0.95857 0.95920 0.95981 0.9604 0.9610

PC147 PC148 PC149 PC150 PC151 PC152 PC153

Standard deviation 1.48538 1.47245 1.46502 1.45808 1.45233 1.44902 1.44139

Proportion of Variance 0.00059 0.00058 0.00058 0.00057 0.00057 0.00056 0.00056

Cumulative Proportion 0.96160 0.96218 0.96276 0.96333 0.96389 0.96446 0.96501

PC154 PC155 PC156 PC157 PC158 PC159 PC160

Standard deviation 1.43556 1.41926 1.41331 1.40976 1.39176 1.38657 1.37934

Proportion of Variance 0.00055 0.00054 0.00054 0.00053 0.00052 0.00052 0.00051

Cumulative Proportion 0.96557 0.96611 0.96664 0.96717 0.96769 0.96821 0.96872

PC161 PC162 PC163 PC164 PC165 PC166 PC167

Standard deviation 1.37514 1.3621 1.35286 1.34895 1.33738 1.33102 1.32491

Proportion of Variance 0.00051 0.0005 0.00049 0.00049 0.00048 0.00047 0.00047

Cumulative Proportion 0.96922 0.9697 0.97021 0.97070 0.97118 0.97165 0.97212

PC168 PC169 PC170 PC171 PC172 PC173 PC174

Standard deviation 1.32213 1.31056 1.30125 1.29600 1.29157 1.28034 1.27200

Proportion of Variance 0.00047 0.00046 0.00045 0.00045 0.00045 0.00044 0.00043

Cumulative Proportion 0.97259 0.97305 0.97351 0.97396 0.97440 0.97484 0.97528

PC175 PC176 PC177 PC178 PC179 PC180 PC181

Standard deviation 1.26228 1.25502 1.24960 1.23701 1.23398 1.23058 1.2218

Proportion of Variance 0.00043 0.00042 0.00042 0.00041 0.00041 0.00041 0.0004

Cumulative Proportion 0.97570 0.97613 0.97655 0.97696 0.97736 0.97777 0.9782

PC182 PC183 PC184 PC185 PC186 PC187 PC188

Standard deviation 1.2164 1.2152 1.20075 1.19609 1.18407 1.18060 1.17167

Proportion of Variance 0.0004 0.0004 0.00039 0.00038 0.00038 0.00037 0.00037

Cumulative Proportion 0.9786 0.9790 0.97935 0.97973 0.98011 0.98048 0.98085

PC189 PC190 PC191 PC192 PC193 PC194 PC195

Standard deviation 1.16575 1.15991 1.15541 1.14914 1.14320 1.12482 1.12045

Proportion of Variance 0.00036 0.00036 0.00036 0.00035 0.00035 0.00034 0.00034

Cumulative Proportion 0.98121 0.98157 0.98193 0.98229 0.98264 0.98298 0.98331

PC196 PC197 PC198 PC199 PC200 PC201 PC202

Standard deviation 1.11027 1.10223 1.09017 1.08168 1.08118 1.07166 1.0657

Proportion of Variance 0.00033 0.00033 0.00032 0.00031 0.00031 0.00031 0.0003

Cumulative Proportion 0.98364 0.98397 0.98429 0.98460 0.98491 0.98522 0.9855

PC203 PC204 PC205 PC206 PC207 PC208 PC209

Standard deviation 1.0549 1.04734 1.03963 1.03364 1.03115 1.01944 1.01487

Proportion of Variance 0.0003 0.00029 0.00029 0.00029 0.00028 0.00028 0.00028

Cumulative Proportion 0.9858 0.98612 0.98641 0.98669 0.98698 0.98726 0.98753

PC210 PC211 PC212 PC213 PC214 PC215 PC216

Standard deviation 1.01137 1.00041 0.99504 0.98119 0.97557 0.97156 0.96250

Proportion of Variance 0.00027 0.00027 0.00027 0.00026 0.00026 0.00025 0.00025

Cumulative Proportion 0.98781 0.98808 0.98834 0.98860 0.98885 0.98911 0.98936

PC217 PC218 PC219 PC220 PC221 PC222 PC223

Standard deviation 0.95918 0.95292 0.94665 0.93992 0.93952 0.93135 0.92426

Proportion of Variance 0.00025 0.00024 0.00024 0.00024 0.00024 0.00023 0.00023

Cumulative Proportion 0.98960 0.98985 0.99009 0.99032 0.99056 0.99079 0.99102

PC224 PC225 PC226 PC227 PC228 PC229 PC230

Standard deviation 0.92057 0.91039 0.90505 0.89710 0.88729 0.87956 0.8741

Proportion of Variance 0.00023 0.00022 0.00022 0.00022 0.00021 0.00021 0.0002

Cumulative Proportion 0.99125 0.99147 0.99169 0.99190 0.99212 0.99232 0.9925

PC231 PC232 PC233 PC234 PC235 PC236 PC237

Standard deviation 0.8606 0.8582 0.8570 0.84249 0.83673 0.82987 0.82862

Proportion of Variance 0.0002 0.0002 0.0002 0.00019 0.00019 0.00018 0.00018

Cumulative Proportion 0.9927 0.9929 0.9931 0.99331 0.99350 0.99368 0.99387

PC238 PC239 PC240 PC241 PC242 PC243 PC244

Standard deviation 0.82508 0.81531 0.80633 0.80084 0.78484 0.77892 0.77493

Proportion of Variance 0.00018 0.00018 0.00017 0.00017 0.00017 0.00016 0.00016

Cumulative Proportion 0.99405 0.99423 0.99440 0.99457 0.99474 0.99490 0.99506

PC245 PC246 PC247 PC248 PC249 PC250 PC251

Standard deviation 0.76113 0.75609 0.75120 0.74252 0.73792 0.73228 0.71553

Proportion of Variance 0.00016 0.00015 0.00015 0.00015 0.00015 0.00014 0.00014

Cumulative Proportion 0.99522 0.99537 0.99552 0.99567 0.99582 0.99596 0.99610

PC252 PC253 PC254 PC255 PC256 PC257 PC258

Standard deviation 0.71204 0.70665 0.69911 0.68803 0.67874 0.66695 0.65021

Proportion of Variance 0.00014 0.00013 0.00013 0.00013 0.00012 0.00012 0.00011

Cumulative Proportion 0.99623 0.99637 0.99650 0.99662 0.99675 0.99687 0.99698

PC259 PC260 PC261 PC262 PC263 PC264 PC265

Standard deviation 0.63331 0.6201 0.6154 0.6046 0.5980 0.58985 0.58688

Proportion of Variance 0.00011 0.0001 0.0001 0.0001 0.0001 0.00009 0.00009

Cumulative Proportion 0.99709 0.9972 0.9973 0.9974 0.9975 0.99758 0.99767

PC266 PC267 PC268 PC269 PC270 PC271 PC272

Standard deviation 0.58060 0.57940 0.56895 0.55909 0.55821 0.55256 0.54885

Proportion of Variance 0.00009 0.00009 0.00009 0.00008 0.00008 0.00008 0.00008

Cumulative Proportion 0.99776 0.99785 0.99794 0.99802 0.99811 0.99819 0.99827

PC273 PC274 PC275 PC276 PC277 PC278 PC279

Standard deviation 0.53891 0.52102 0.51198 0.50837 0.50286 0.49930 0.48972

Proportion of Variance 0.00008 0.00007 0.00007 0.00007 0.00007 0.00007 0.00006

Cumulative Proportion 0.99835 0.99842 0.99849 0.99856 0.99863 0.99869 0.99876

PC280 PC281 PC282 PC283 PC284 PC285 PC286

Standard deviation 0.48420 0.47491 0.46108 0.45560 0.44980 0.43244 0.42183

Proportion of Variance 0.00006 0.00006 0.00006 0.00006 0.00005 0.00005 0.00005

Cumulative Proportion 0.99882 0.99888 0.99894 0.99899 0.99905 0.99910 0.99915

PC287 PC288 PC289 PC290 PC291 PC292 PC293

Standard deviation 0.41475 0.40910 0.40551 0.39753 0.39449 0.38137 0.37235

Proportion of Variance 0.00005 0.00004 0.00004 0.00004 0.00004 0.00004 0.00004

Cumulative Proportion 0.99919 0.99924 0.99928 0.99932 0.99936 0.99940 0.99944

PC294 PC295 PC296 PC297 PC298 PC299 PC300

Standard deviation 0.36933 0.36672 0.33968 0.33654 0.33337 0.33019 0.32799

Proportion of Variance 0.00004 0.00004 0.00003 0.00003 0.00003 0.00003 0.00003

Cumulative Proportion 0.99948 0.99951 0.99954 0.99957 0.99960 0.99963 0.99966

PC301 PC302 PC303 PC304 PC305 PC306 PC307

Standard deviation 0.32321 0.31775 0.30176 0.29443 0.28287 0.27161 0.25923

Proportion of Variance 0.00003 0.00003 0.00002 0.00002 0.00002 0.00002 0.00002

Cumulative Proportion 0.99969 0.99972 0.99974 0.99977 0.99979 0.99981 0.99982

PC308 PC309 PC310 PC311 PC312 PC313 PC314

Standard deviation 0.25177 0.24988 0.23687 0.22990 0.21634 0.21479 0.20454

Proportion of Variance 0.00002 0.00002 0.00002 0.00001 0.00001 0.00001 0.00001

Cumulative Proportion 0.99984 0.99986 0.99987 0.99989 0.99990 0.99991 0.99992

PC315 PC316 PC317 PC318 PC319 PC320 PC321

Standard deviation 0.19120 0.18473 0.18353 0.17965 0.16214 0.1333 0.1302

Proportion of Variance 0.00001 0.00001 0.00001 0.00001 0.00001 0.0000 0.0000

Cumulative Proportion 0.99993 0.99994 0.99995 0.99996 0.99997 1.0000 1.0000

PC322 PC323 PC324 PC325 PC326 PC327 PC328

Standard deviation 0.1259 0.1184 0.114 0.101 0.09562 0.09234 0.08783

Proportion of Variance 0.0000 0.0000 0.000 0.000 0.00000 0.00000 0.00000

Cumulative Proportion 1.0000 1.0000 1.000 1.000 0.99999 1.00000 1.00000

PC329 PC330 PC331 PC332 PC333 PC334

Standard deviation 0.05892 0.05013 0.04424 0.02842 1.531e-13 2.615e-14

Proportion of Variance 0.00000 0.00000 0.00000 0.00000 0.000e+00 0.000e+00

Cumulative Proportion 1.00000 1.00000 1.00000 1.00000 1.000e+00 1.000e+00

PC335 PC336 PC337 PC338 PC339 PC340

Standard deviation 8.496e-15 3.02e-15 3.02e-15 3.02e-15 3.02e-15 3.02e-15

Proportion of Variance 0.000e+00 0.00e+00 0.00e+00 0.00e+00 0.00e+00 0.00e+00

Cumulative Proportion 1.000e+00 1.00e+00 1.00e+00 1.00e+00 1.00e+00 1.00e+00

PC341 PC342 PC343 PC344 PC345 PC346

Standard deviation 3.02e-15 3.02e-15 3.02e-15 3.02e-15 3.02e-15 3.02e-15

Proportion of Variance 0.00e+00 0.00e+00 0.00e+00 0.00e+00 0.00e+00 0.00e+00

Cumulative Proportion 1.00e+00 1.00e+00 1.00e+00 1.00e+00 1.00e+00 1.00e+00

PC347 PC348 PC349 PC350 PC351 PC352

Standard deviation 3.02e-15 3.02e-15 3.02e-15 3.02e-15 3.02e-15 3.02e-15

Proportion of Variance 0.00e+00 0.00e+00 0.00e+00 0.00e+00 0.00e+00 0.00e+00

Cumulative Proportion 1.00e+00 1.00e+00 1.00e+00 1.00e+00 1.00e+00 1.00e+00

PC353 PC354 PC355 PC356 PC357 PC358

Standard deviation 3.02e-15 3.02e-15 3.02e-15 3.02e-15 3.02e-15 3.02e-15

Proportion of Variance 0.00e+00 0.00e+00 0.00e+00 0.00e+00 0.00e+00 0.00e+00

Cumulative Proportion 1.00e+00 1.00e+00 1.00e+00 1.00e+00 1.00e+00 1.00e+00

PC359 PC360 PC361 PC362 PC363 PC364

Standard deviation 3.02e-15 3.02e-15 3.02e-15 3.02e-15 3.02e-15 3.02e-15

Proportion of Variance 0.00e+00 0.00e+00 0.00e+00 0.00e+00 0.00e+00 0.00e+00

Cumulative Proportion 1.00e+00 1.00e+00 1.00e+00 1.00e+00 1.00e+00 1.00e+00

PC365 PC366 PC367 PC368 PC369 PC370

Standard deviation 3.02e-15 3.02e-15 3.02e-15 3.02e-15 3.02e-15 3.02e-15

Proportion of Variance 0.00e+00 0.00e+00 0.00e+00 0.00e+00 0.00e+00 0.00e+00

Cumulative Proportion 1.00e+00 1.00e+00 1.00e+00 1.00e+00 1.00e+00 1.00e+00

PC371

Standard deviation 1.538e-15

Proportion of Variance 0.000e+00

Cumulative Proportion 1.000e+00

>my\_data<- subset(tISmatrixZooInGenome, select = -c(Result))

> test.data1 <- predict(pca1, newdata = my\_data)

> test.data1 <- as.data.frame(test.data1)

> test.data1 <- data.frame(Result = tISmatrixZooInGenome$Result, test.data1)

Error in tISmatrixZooInGenome$Result :

$ operator is invalid for atomic vectors

> tISmatrixZooInGenome<- as.data.frame( tISmatrixZooInGenome)

> test.data1 <- data.frame(Result = tISmatrixZooInGenome$Result, test.data1)

Warning message:

In data.row.names(row.names, rowsi, i) :

some row.names duplicated: 3,4,5 --> row.names NOT used

> test.data1 <- test.data1[,1:130]

> write.csv(test.data1, file = "C:/Ava/android/Trace/Zoo2InGenome.csv")

///////////////////////////////////////////

Error matrix for the Decision Tree model on Zoo2Googleplay2.csv [test] (counts):

Predicted

Actual 0 1

0 25 4

1 6 22

Error matrix for the Decision Tree model on Zoo2Googleplay2.csv [test] (proportions):

Predicted

Actual 0 1 Error

0 0.44 0.07 0.14

1 0.11 0.39 0.21

Overall error: 18%, Averaged class error: 18%

Rattle timestamp: 2017-08-22 12:37:26

======================================================================

Error matrix for the Random Forest model on Zoo2Googleplay2.csv [test] (counts):

Predicted

Actual 0 1

0 27 2

1 3 25

Error matrix for the Random Forest model on Zoo2Googleplay2.csv [test] (proportions):

Predicted

Actual 0 1 Error

0 0.47 0.04 0.07

1 0.05 0.44 0.11

Overall error: 9%, Averaged class error: 9%

Rattle timestamp: 2017-08-22 12:37:26

======================================================================

Error matrix for the SVM model on Zoo2Googleplay2.csv [test] (counts):

Predicted

Actual 0 1

0 10 19

1 3 25

Error matrix for the SVM model on Zoo2Googleplay2.csv [test] (proportions):

Predicted

Actual 0 1 Error

0 0.18 0.33 0.66

1 0.05 0.44 0.11

Overall error: 39%, Averaged class error: 38%

Rattle timestamp: 2017-08-22 12:37:26

======================================================================

Area under the ROC curve for the rpart model on Zoo2Googleplay2.csv [test] is 0.8233

Rattle timestamp: 2017-08-22 12:49:05

======================================================================

Area under the ROC curve for the rf model on Zoo2Googleplay2.csv [test] is 0.9255

Rattle timestamp: 2017-08-22 12:49:05

======================================================================

Area under the ROC curve for the ksvm model on Zoo2Googleplay2.csv [test] is 0.8017

Rattle timestamp: 2017-08-22 12:49:05

======================================================================