Enhancement of Activity Recognition (AR) Performance over Less Frequent Activies in a Non-Uniform Dataset

By: Mohsen Nabian

PhD student in MIE department at Northeastern University

## Abstract:

Active Recognition (AR) systems are developed to recognize human activities in real-time bases in cellphones and smartwatches. These techniques, are widely used in many mobile applications to recognize the user's real-time activity like walking,sitting, running and biking.

### Q1) Summerize what is done sofar:

In this paper, we implemented several classification algorithms on a AR dataset[[1]](#footnote-1). The goal of this research are the following:

1. comparing the predictive ability of the following classifiers on AR data: -SVM -KNN -Artificial Neural Network -Decision Tree -Bagging -Rndom Forest -Naive Base
2. The data set is highly biased on one specific activity (sitting). This non-uniformity must be addressed to improve predictions. By modifying of the training data aiming to reduce non-uniformity, we discovered the prediction is improved for the other activities with the cost of less accurate prediction for the sitting activity. The goal is to find an optimum point to improve the prediction.
3. Finally, we would also compare our classifier accuracy results with state-of-the-art classifiers for Activity Recognition like FE-AT2.

### Q2) Quick discussion on why this research is intresting.

This research provides a brief understanding of how the most famous classifiers would perform over the same training data as provided. Moreover, It indicates how traing data modification can affect the learning algorithm performance. Furthemore, It provides a final reliable classifier and compares it to the newly introduced classifiers called FE-AT.

### Q3) What has been done sofar:

So far, the Data has been extracted, cleaned, normalized and shuffled.In addition, an exploratory statistics proceadure is to be done over the features and labels.Then 6 classifiers have been used for classification and their performance were measured based on the correct prediction for the dominant activity (sitting) as well as the rare activities. Using the current dataset, leads to very accurate prediction of the dominant activity, however, it is unsuccessful to predict rare activities (76%). We intend to increase this accuracy by manupulationg the training set. ###Reading the data

dt1<-read.table("/Users/mohsennabian/Desktop/MHEALTHDATASET/mHealth\_subject1.log")  
summary(dt1)  
dim(dt1)  
  
dt2<-read.table("/Users/mohsennabian/Desktop/MHEALTHDATASET/mHealth\_subject2.log")  
summary(dt2)  
dim(dt2)  
  
dt3<-read.table("/Users/mohsennabian/Desktop/MHEALTHDATASET/mHealth\_subject3.log")  
summary(dt3)  
dim(dt3)  
  
dt4<-read.table("/Users/mohsennabian/Desktop/MHEALTHDATASET/mHealth\_subject4.log")  
summary(dt4)  
dim(dt4)  
  
dt5<-read.table("/Users/mohsennabian/Desktop/MHEALTHDATASET/mHealth\_subject5.log")  
summary(dt5)  
dim(dt5)  
  
dt6<-read.table("/Users/mohsennabian/Desktop/MHEALTHDATASET/mHealth\_subject6.log")  
summary(dt6)  
dim(dt6)  
  
dt7<-read.table("/Users/mohsennabian/Desktop/MHEALTHDATASET/mHealth\_subject7.log")  
summary(dt7)  
dim(dt7)  
  
dt8<-read.table("/Users/mohsennabian/Desktop/MHEALTHDATASET/mHealth\_subject8.log")  
summary(dt8)  
dim(dt8)  
  
dt9<-read.table("/Users/mohsennabian/Desktop/MHEALTHDATASET/mHealth\_subject9.log")  
summary(dt9)  
dim(dt9)  
  
dt10<-read.table("/Users/mohsennabian/Desktop/MHEALTHDATASET/mHealth\_subject10.log")  
summary(dt10)  
dim(dt10)  
  
all<-rbind(dt1,dt2,dt3,dt4,dt5,dt6,dt7,dt8,dt9,dt10)  
  
  
names(all)[24]<-"Y"  
names(all)  
  
all$Y<-as.factor(all$Y)  
summary(all)  
str(all)  
  
# So 'all' is the all the data we have. 23 attributes (column 1-23) and 1 label(column 24) with 13 classes.

normalizing data:

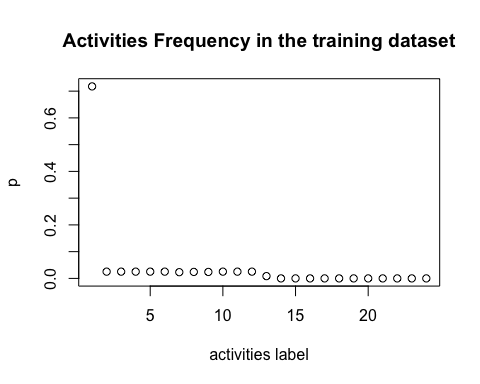
library('som')  
all\_norm<-normalize(all[,-24], byrow=FALSE)  
all\_norm<-as.data.frame(all\_norm)  
all\_norm$Y<-all$Y  
head(all\_norm)  
  
#'all\_norm' is the 'all' data frame but normalized.   
  
  
dt<-all\_norm[sample(nrow(all\_norm), nrow(all\_norm)), ]  
  
#dt is the shuffled format of 'all\_norm'

data explanatory:

a=c()  
for (i in 1:(ncol(dt)-1))  
{  
a[i]<-sum(dt$Y==i)/nrow(dt)  
}  
  
# a is the frequency vector of each label from a[1] to a[12]  
  
b<-sum(dt$Y==0)/nrow(dt)  
  
#b is the frequency of label 0  
  
plot(a)

As shown above, the data is hetrogenous and is highly biased toward the 1st label which is sitting.

cc<-c(b,a)  
#cc is a frequency label of all Y classes cc[1]..cc[13]  
plot(cc,ylab="p",xlab="activities label",main="Activities Frequency in the training dataset")



Train and Test data

ntr<-30000 #training  
nte<-1000 #test  
  
r<-0.8 #percentage of training data to be selected from rare activities.

Method I)

train1<-dt[1:ntr,]  
dim(train1)  
  
test1<-dt[(ntr+1):(ntr+nte),]  
dim(test1)

some evaluation functions:

true\_rare<-function(test,ypred)  
{  
 r<-length(test$Y)-sum(test$Y==0)  
 r  
 a<-sum((ypred==test$Y) & (test$Y!=0) )/r  
 return(a)  
}  
  
true\_zero<-function(test,ypred)  
{  
 r<-sum(test$Y==0)  
 a<-sum((ypred==test$Y) & (test$Y==0) )/r  
 return(a)  
}

### SVM

library(e1071)   
svm.fit1<- svm(Y~.,data=train1, kernal="polynomial",cost=100,scale=TRUE)  
ypred\_svm1=predict(svm.fit1,test1)  
  
true\_rare(test1,ypred\_svm1)  
true\_zero(test1,ypred\_svm1)

### KNN

library(class)  
ypred\_knn1<-knn(train1[,-24], test1[,-24], train1[,24], k = 13, prob=FALSE)  
ypred\_knn2<-knn(train2[,-24], test2[,-24], train2[,24], k = 13, prob=FALSE)  
  
#Evaluation:  
  
true\_rare(test1,ypred\_knn1)  
true\_zero(test1,ypred\_knn1)

### Artificial Neural Network

library("nnet")  
#Changing Y label to binary labels:  
ideal1<-as.data.frame(model.matrix(~0+train1[,24])) #or ideal <- class.ind(train[,24])  
seedsANN = nnet(train1[,-24],ideal1, size=20, softmax=TRUE,maxit = 200)  
ypred\_nnet1<-predict(seedsANN, test1[,-24],type="class")  
#ypred\_nnet1 is in the string format, need to be modified:  
library(stringr)  
ypred\_nnet1<-as.numeric(str\_extract(ypred\_nnet1,"[0-9][0-9]\*$"))  
#Evaluation:  
  
true\_rare(test1,ypred\_nnet1)  
true\_zero(test1,ypred\_nnet1)

### Decision Tree

library('C50')  
  
set.seed(12345)  
model1 <- C5.0(train1[,-24], train1[,24])  
ypred\_dt1<-predict(model1, test1[,-24])  
  
#Evaluation:  
  
true\_rare(test1,ypred\_dt1)  
true\_zero(test1,ypred\_dt1)

### Bagging

# load the package  
library(ipred)  
  
fit1 <- bagging(Y~., data=train1)  
ypred\_bg1<- predict(fit1, test1[,-24], type="class")  
  
#Evaluation:  
  
true\_rare(test1,ypred\_bg1)  
true\_zero(test1,ypred\_bg1)

### random forest

library(randomForest)  
fit1<- randomForest(Y~., data=train1)  
ypred\_rf1<- predict(fit1, test1[,-24], type="class")  
#Evaluation:  
  
true\_rare(test1,ypred\_rf1)  
true\_zero(test1,ypred\_rf1)

### Naieve-Base

library('e1071')  
model1 <- naiveBayes(Y~., data=train1)  
ypred\_naivbs1<-predict(model1, test1[,-24])  
#Evaluation:  
  
true\_rare(test1,ypred\_naivbs1)  
true\_zero(test1,ypred\_naivbs1)

The predication results from classifiers are summerized and demonstrated in the following figure:

Fig1. predication accuracy for Activity Recognition problem using a heterogeneous dataset. Blue and Red bars represent True-Positive rare activities prediction and True-Positive dominant activity respectively.

### Q4) What is the next step?

The Next steps are to modify the training set as well as the classifiers variables with the goal of improving the overal predictive performance of both the dominant activity (sitting) as well as rare activities. Moreover, the performance is to be compared with the newest algorithms in Activivty Recognition problems.Finally, a new algorithm might be introduced to improve the current most accurate predictions and lead to a contribution in this field.

1. UCI Machine Learning Dataset

   2 Recognizing New Activities with Limited Training Data , JANUARY 2015 DOI: 10.1145/2802083.2808388 [↑](#footnote-ref-1)