Table of Contents

Pre-Processing DataSet

```
function [final_feature,variables,label,rowdh] =
 preprocessing(feature, variables,label,tt)
```

Recasting the representation of features

```
pp_feature = zeros(size(feature,1),size(feature,2));
final_feature = zeros(size(feature,1),size(feature,2));
for i = 1 : size(feature,2)
    if (iscellstr(table2cell(feature(1,i))) == 1)
        uni = unique(feature(:,i));
        for j = 1 : size(unique(feature(:,i)))
            if i >= 25 && i <=47
                uni = cell2table([{'No'};{'Down'};{'Steady'};{'Up'}]);
            end
            pp_feature(:,i) = j *
 strcmp(table2cell(feature(:,i)),table2cell(uni(j,1)));
            final_feature(:,i) = final_feature(:,i) + pp_feature(:,i);
        end
    else
        final_feature(:,i) = cell2mat(table2cell(feature(:,i)));
    end
end
```

Removing NaN and replacing them with mode values

```
[row, ~] = find(isnan(final_feature(:,19)));
final_feature(row(:),:) = []; label(row(:),:) = [];
% KNN Imputaiton
% feature_train_pp(:,19:21) = knnimpute(feature_train_pp(:,19:21));
```

Removing Sample points

Duplicate Patients

```
[~,unind,~] = unique(final_feature(:,2),'rows','stable');
final_feature = final_feature(unind(:),:); label = label(unind(:),:);
% Removal of dead or hospice patients
rowdh = find(ismember(final_feature(:,8),[11 13 14 19 20 21 8 18 25
 26]));
final_feature(rowdh(:),:) = []; label(rowdh(:),:) = [];
```

Removal of invalid genders

```
rowg = find(ismember(final_feature(:,4),3));
final_feature(rowg(:),:) = []; label(rowg(:),:) = [];
```

Histogram Merging

final_feature(find(ismember(final_feature(:,3),[3 5 6])),3) = 1;

```
% final_feature(find(ismember(final_feature(:,7),[4:9])),7) = 4;
% final_feature(find(ismember(final_feature(:,8),[25 26])),8) = 4;
% final_feature(find(ismember(final_feature(:,8),[6])),8) = 5;
% final_feature(find(ismember(final_feature(:,9),[2:6 8:16
 18:end])),9) = 2;
% final_feature(find(ismember(final_feature(:,9),[7])),9) = 3;
% final_feature(find(ismember(final_feature(:,9),[17])),9) = 4;
```

Merging data points/categories in the same feature space

```
final_feature(find(ismember(final_feature(:,3),[3 5 6])),3) = 3;

final_feature(find(ismember(final_feature(:,5),[2:5])),5) = 1;
final_feature(find(ismember(final_feature(:,5),[6])),5) = 2;
final_feature(find(ismember(final_feature(:,5),[7])),5) = 3;
final_feature(find(ismember(final_feature(:,5),[8])),5) = 4;
final_feature(find(ismember(final_feature(:,5),[9:10])),5) = 5;

final_feature(find(ismember(final_feature(:,7),[1 2 7])),7) = 1;
final_feature(find(ismember(final_feature(:,7),[3 4])),7) = 2;
final_feature(find(ismember(final_feature(:,7),[5 6 8])),7) = 3;

final_feature(find(ismember(final_feature(:,8),[1 6 8])),8) = 1;
final_feature(find(ismember(final_feature(:,8),[2:30])),8) = 2;

final_feature(find(ismember(final_feature(:,9), [2 3])),9) = 1;
final_feature(find(ismember(final_feature(:,9), [7])),9) = 2;
final_feature(find(ismember(final_feature(:,9),[3:26])),9) = 3;

final_feature(find(ismember(final_feature(:,12),[2:4 7:11 13:55
 69])),12) = 2;
final_feature(find(ismember(final_feature(:,12),[5 6])),12) = 3;
final_feature(find(ismember(final_feature(:,12),[12])),12) = 4;
final_feature(find(ismember(final_feature(:,12),[19])),12) = 5;
final_feature(find(ismember(final_feature(:,12),[56:68])),12) = 6;
```

ICD9 Mapping/Merging

```
icd9 = [0 139 239 279 289 319 359 389 459 519 579 629 679 709 739 759
 779 799 999 1000 2000];
final_feature(find(fix(final_feature(:,19)) ==  250),i)= 2000;
final_feature(find(fix(final_feature(:,19)) ~=
 final_feature(:,19)),19)= 1000;
final_feature(find(ismember(final_feature(:,19),[780 781 784 790:799
 240:279 680:709 782 ...
    1:139 1000 280:289 320:359 630:679 360:389 740:759])),19) = 1;
final_feature(find(ismember(final_feature(:,19),[390:450 785])),19) =
 2;
final_feature(find(ismember(final_feature(:,19),[460:519 786])),19) =
 3;
final_feature(find(ismember(final_feature(:,19),[520:579 787])),19) =
 4;
final_feature(find(ismember(final_feature(:,19),[2000])),19) = 5;
final_feature(find(ismember(final_feature(:,19),[800:999])),19) = 6;
final_feature(find(ismember(final_feature(:,19),[710:739])),19) = 7;
final_feature(find(ismember(final_feature(:,19),[580:629 788])),19) =
 8;
final_feature(find(ismember(final_feature(:,19),[140:239])),19) = 9;

rowdg = find(ismember(final_feature(:,19),[10:2000]));
final_feature(rowdg(:),:) = []; label(rowdg(:),:) = [];
```

Feature Demensionality Reduction

```
correl = corrcoef(final_feature,'rows','pairwise');
fdr = [1 2 6 11 20 21 23:41 43:47]; %38 40 41 45];%
final_feature(:,fdr(:)) = [];   variables(:,fdr(:)) = [];
```

Categorical Feature Exapnsion

```
for h = [1:6 8 15]
    final_feature(:,h) = 2.^(final_feature(:,h)-1);
    bin = dec2bin(final_feature(:,h));
    sz = size(final_feature,2);
    for i = 1 : size(unique(final_feature(:,h)),1)
        final_feature(:,sz+i) = bin(:,i)-48;
    end
end
final_feature(:,[1:6 8 15]) = []; variables(:,[1:6 8 15]) = [];
```

```
vari = var(final_feature); % Find near to zero variance features and
 remove them
final_feature(:,find(vari<0.1)) = [];
```

Outlier Detection and Removal

```
[~,idx] = outliers(final_feature,1000);
rowout = idx(:,1);
final_feature(rowout(:),:) = []; label(rowout(:),:) = [];
```

Normalizing Data

```
final_feature = zscore(final_feature); % Normalize the features
```

Unique Data Set Required

```
[~,unind,~] = unique(final_feature(:,:),'rows','stable');
final_feature = final_feature(unind(:),:); label = label(unind(:),:);
```

Published with MATLAB® R2016a