Dimensionality Reduction via Dynamic Clustering and Stratified Random Sampling (DCSRandS)

Shuman Peng
Simon Fraser University
November 20, 2019
Overview

- Motivation & Problem Setting
- Problem Definition
- Existing Methods
- Proposed Method - DCSRandS
- Experimental Evaluation
- Conclusion
- Future Works
Motivation

● Motivating Application
  ○ Pharmacogenomics
    ■ **Goal**: predict drug response in patients using their genomic information (crucial task of *precision oncology*)

● Characteristics of the data
  ○ Cisplatin: 118 samples (patients) x 11768 features (genes)
    - Train: 78, Test: 40
    - 94 S (1) R (0)
  ○ Bortezomib: 236 samples (patients) x 11609 features (genes)
    - Train: 156, Test: 80
    - 124 S (1) R (0)

Lack of data & High dimensionality
Problem Definition

- **Given:**
  - A pharmacogenomics dataset $D$ that consists of very few observations and a very large number of features
  - Train a classification model that can accurately predict the drug response

- **Challenges:**
  - Very high feature dimension!
  - Overfitting / poor prediction accuracy

- **Solution:**
  - Use sampling methods to select meaningful features to reduce the dimensionality of the data
  - Simple and efficient
Possible Solutions

1. Method 1: Reduce the scope of the original problem (SRSWOR)
   - Pros:
     - Easy & fast
     - Reduces the dimension
   - Cons:
     - The dimension can still be high
     - Information loss

2. Method 2: Further reduce the dimensionality by SRSWOR on the reduced feature scope
   - Pros:
     - Lower dimensionality (fewer features)
   - Cons:
     - Information loss → are the remaining features meaningful?
Possible Solutions

3. Method 3: Stratified simple random sampling on clusters from K-Means
   ○ Pros:
     ■ Lower dimensionality
     ■ More “meaningful” features → removes redundancy & solves overfitting
   ○ Cons:
     ■ Euclidean distance is used → how accurately can this measure the similarity between features?
     ■ K-Mean clustering requires the number of clusters to be specified

 Can we create more meaningful clusters and do it efficiently?
Proposed Method

- **Dynamic Clustering and Stratified Random Sampling (DCSRandS)**
- Can effectively reduce the dimensionality of the data while selecting meaningful features (removes redundancy without losing useful information)

- Consists of 2 major components:
  - **Dynamic Clustering** based on feature correlation (strata creation)
    - Meaningful clusters → do not need to specify the number of clusters
  - **Stratified Random Sampling**
    - Removes redundancy

- Limitations:
  - Needs to pre-compute an \( n \times n \) correlation matrix for every pair of features in order to perform Dynamic Clustering
  - Computationally expensive when \( n \) is very large (~ 10,000)
    - Reduce the scope of the original problem
DCSRandS Overview

Very large $n$ (> 10,000)

<table>
<thead>
<tr>
<th>$f_1$</th>
<th>$f_2$</th>
<th>...</th>
<th>$f_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$l$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

# of clusters unknown ahead of time

Dynamic Feature Clustering

Stratified Random Sampling

Predict the response 0 / 1

SVM Ensemble (k SVMs)

$C \equiv \# \text{ of clusters}$

$k = 3$

$\{f_{i1}, f_{i2}, ..., f_{ic}\}$

$\{f_{21}, f_{22}, ..., f_{2c}\}$

$\vdots$

$\{f_{k1}, f_{k2}, ..., f_{kc}\}$
Motivation:

○ Existing clustering methods (i.e., K-Means) require the number of clusters to be specified, but the "ideal" number of clusters is unknown ahead of time and requires guessing and testing.

○ It also does not make sense to "force" a feature to be grouped into any of the existing clusters when it should be in its own separate cluster.

○ Solution: dynamic clustering. Create new clusters (when necessarily) as you go.

How is Dynamic Clustering done?

○ Cluster any un-clustered feature $f_i$ into (1) one of the existing clusters, say $C_j$, if the Pearson Correlation Coefficient between $f_i$ and every feature $f \in C_j$ exceeds a minimum correlation threshold $\tau$, or (2) create a new cluster $C_p$ and assign $f_i$ to $C_p$ if $f_i$ does not meet the minimum correlation requirement with all the features in any of the existing clusters.
DCSRandS: Stratified Random Sampling

● Motivation:
  ○ We want to reduce the dimensionality of the data by removing redundant features and preserving useful information.
  ○ Solution: Stratified Random Sampling on the feature clusters computed by Dynamic Clustering.

● How is Stratified Random Sampling (SRS) done?
  1. Randomly sample one feature (with replacement) from each feature stratum (cluster) and add it to the set of sampled features.
  2. Repeat step 1) \(k\) times, each time creating a new set of sampled features. At the end, there should be \(k\) sets of samples resulting from this SRS procedure.
Experimental Design

- Experimental Questions:
  1. Can the proposed method effectively reduce the dimensionality of the data?
  2. Can the proposed method improve the performance of the classifier compared to existing methods?

- Baseline methods:
  1. No sampling – using all the features (~ 10,000 features)
  2. One round of sampling (SRSWOR) to reduce the scope of the problem
  3. Simple random sampling without replacement on (2)
  4. Stratified random sampling on K-Means feature clusters on (2)

- Performance measures: AUROC, AUPR
- Classifier: SVM
### Experimental Results - Cisplatin

<table>
<thead>
<tr>
<th></th>
<th>DCSRandS</th>
<th>Baseline 1 (all features)</th>
<th>Baseline 2 (reduced scope (RS))</th>
<th>Baseline 3 (SRSWOR on RS)</th>
<th>Baseline 4 (SRS – KMeans)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AUROC</strong></td>
<td>0.55 (0.07)</td>
<td>0.49 (0.04)</td>
<td>0.49 (0.06)</td>
<td>0.48 (0.07)</td>
<td>0.5 (0.0)</td>
</tr>
<tr>
<td><strong>AUPR</strong></td>
<td>0.81 (0.03)</td>
<td>0.79 (0.01)</td>
<td>0.79 (0.02)</td>
<td>0.79 (0.02)</td>
<td>0.80 (0.002)</td>
</tr>
</tbody>
</table>

- The AUPR (average precision) is high due to the imbalanced data (much more data in the positive class)
- DSRandS leads to improved classification performance
## Experimental Results - Bortezomib

<table>
<thead>
<tr>
<th></th>
<th>DCSRandS</th>
<th>Baseline 1 (all features)</th>
<th>Baseline 2 (reduced scope (RS))</th>
<th>Baseline 3 (SRSWOR on RS)</th>
<th>Baseline 4 (SRS – KMeans)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AUROC</strong></td>
<td>0.63 (0.06)</td>
<td>0.63 (0.01)</td>
<td>0.57 (0.04)</td>
<td>0.60 (0.02)</td>
<td>0.56 (0.02)</td>
</tr>
<tr>
<td><strong>AUPR</strong></td>
<td>0.61 (0.04)</td>
<td>0.60 (0.01)</td>
<td>0.57 (0.02)</td>
<td>0.58 (0.02)</td>
<td>0.56 (0.01)</td>
</tr>
</tbody>
</table>

- The performance from using DCSRandS is comparable to the performance of using Baseline 1 (using all the features)
  - DCSRandS was performed with a reduced scope
  - Information loss
- DCSRandS improves the performance of the classifier
Revisiting Experimental Questions

- Experimental Questions:
  1. Can the proposed method effectively reduce the dimensionality of the data?
     - Yes
     - The resulting number of clusters (from DC) is ~ 100 – 200, depending on the dataset
  2. Can the proposed method improve the performance of the classifier compared to existing methods?
     - Yes
Conclusion

- Proposed a method (DCSRandS) that effectively reduces the dimensionality of the data using:
  - Stratified Random Sampling
  - Dynamic Clustering
- Dynamic Clustering based on feature correlation can create more meaningful clusters (strata) compared to K-Means Clustering
- Stratified Random Sampling on computed clusters (strata) removes redundancy
- DCSRandS improves the classification performance
Future Works

- Improve the running time of the $n \times n$ correlation matrix computation algorithm so we can perform DCSRandS on the full feature scope.
- Apply another feature selection / dimensionality reduction method on top of DCSRandS.
Thank You!

Any questions?