Department of Computer Science

Constrained Logistic Regression for Discriminative Pattern Mining

Rajul Anand Chandan Reddy

September 6, 2011

Overview

❖ Overview
Introduction and Motivation
Preliminaries
Proposed Framework
Results
Conclusion

Introduction and Motivation

Preliminaries

Proposed Framework

Results

Conclusion

Introduction

❖ Overview

Introduction and Motivation

❖ Real Life Examples

❖ Challenges

❖ Existing Approaches

❖ Need for Constrained Models

❖ Goals

❖ Contributions

Preliminaries

Proposed Framework

Results

Conclusion

- Differences between subgroups of multivariate dataset is a challenging problem
- We study this problem in context of supervised scenario
- Our emphasis is to highlight the differences between two subgroups of multivariate data while maintaining the class discrimination

Real Life Examples

❖ Overview

Introduction and Motivation

❖ Real Life Examples

❖ Challenges

❖ Existing Approaches

❖ Need for Constrained Models

❖ Goals

❖ Contributions

Preliminaries

Proposed Framework

Results

Conclusion

- Identifying survival behavior of cancer patients across different racial groups spreading across various geographical locations
- Comparing gender discrimination in jobs across different divisions of an organization
- Bias in loan approval to applicants among various branches of banks

Challenges

❖ Overview

Introduction and Motivation

❖ Real Life Examples

❖ Challenges

❖ Existing Approaches

❖ Need for Constrained Models

❖ Goals

❖ Contributions

Preliminaries

Proposed Framework

Results

Conclusion

- Need to understand the kind of changes
- How to detect and model such changes
- Difficult to quantify model based differences between datasets
- More complex the model learning, more tedious to generate comparable models

Existing Approaches

❖ Overview

Introduction and Motivation

❖ Real Life Examples

❖ Challenges

❖ Existing Approaches

❖ Need for Constrained Models

❖ Goals

❖ Contributions

Preliminaries

Proposed Framework

Results

Conclusion

- Prior approaches detected differences in datasets
  - Based on probability distributions between individual attributes (like KL divergence, KS-test)
  - Based on support level of attribute-value combinations (like Contrast sets [2], Subgroup discovery [6], Emerging Pattern mining [4])
- Related change detection [7] and change mining [9] approaches
- Need for an approach that considers the underlying class distribution while estimating the difference between the datasets

Need for Constrained Models

- ❖ Overview
- Introduction and Motivation
- ❖ Real Life Examples
- ❖ Challenges
- ❖ Existing Approaches
- ❖ Need for Constrained Models
- ❖ Goals
- ❖ Contributions
- Preliminaries
- Proposed Framework
- Results
- Conclusion

- Differences in multivariate data distributions based on **model** vary from previous approaches
- Directly obtaining classification models for difference analysis pose questions like
  - Which model can accurately represent the data?
  - Which model to choose among models with similar accuracy?
- Choosing maximum margin classifier model for comparison won’t work
- Number of potential models increase in non-linearly separable case

Goals

❖ Overview

Introduction and Motivation

❖ Real Life Examples

❖ Challenges

❖ Existing Approaches

❖ Need for Constrained Models

❖ Goals

❖ Contributions

Preliminaries

Proposed Framework

Results

Conclusion

- Quantify the change between datasets as the change in underlying class distributions
- Model based class distribution difference instead of data dependent measures
- For the task of discriminative pattern mining
  - The methods for modeling the data should go beyond optimizing a standard prediction metric
  - And should simultaneously identify and model the differences between two multivariate data distributions.

Contributions

❖ Overview
Introduction and Motivation
❖ Real Life Examples
❖ Challenges
❖ Existing Approaches
❖ Need for Constrained Models
❖ Goals
❖ Contributions
Preliminaries
Proposed Framework
Results
Conclusion

- Developed a measure of the distance between two data distributions using the difference between predictive models.
- Developed a constrained version of logistic regression algorithm that can capture the proposed distance measure.
- Experimental justification from results that proposed method quantitatively capture the difference in data distributions

Notations

❖ Overview
Introduction and Motivation
Preliminaries
❖ Notations
❖ Logistic Regression
❖ Logistic Regression Cont'd
❖ Supervised Distribution Difference
Proposed Framework
Results
Conclusion

Notation	***Description***
L	Objective function
C	Regularization factor
w_k	k^{th} component of weight vector w
W_j	j^{th} weight vector
ϵ	Constraint on weight values
μ	Mean
σ	Standard deviation

- **Differential features**:features which are more important in one dataset but less important in the other dataset with respect to classification

Logistic Regression

❖ Overview

Introduction and Motivation

Preliminaries

❖ Notations

❖ Logistic Regression

❖ Logistic Regression Cont'd

❖ Supervised Distribution Difference

Proposed Framework

Results

Conclusion

- Logistic Regression for binary classification
- $\log \frac{\Pr(y=+1|\vec{x})}{\Pr(y=-1|\vec{x})} = \sum_{k=0}^{l} w_k x_k$
- LR learn weights by maximizing the log-likelihood of
- $L(\vec{w}) = \sum_{i=1}^{n} \log \Pr(y = y_i | \vec{x_i}) = \sum_{i=1}^{n} \log g(y_i z_i)$
- Newton's method iteratively updates the weights using the following update equation :
- $\vec{w}^{(t+1)} = \vec{w}^{(t)} - \left[\frac{\partial^2 L}{\partial \vec{w} \partial \vec{w}}\right]^{-1} \frac{\partial L}{\partial \vec{w}}$

Logistic Regression Cont'd

❖ Overview

Introduction and Motivation

Preliminaries

❖ Notations

❖ Logistic Regression

❖ Logistic Regression Cont'd

❖ Supervised Distribution Difference

Proposed Framework

Results

Conclusion

- Final minimization problem with objective function
- $L = -\sum_{i=1}^{n} \log g\left(y_i z_i\right) + \frac{C}{2}\sum_{k=1}^{l} w_k^2$
- $\frac{\partial L}{\partial w_k} = -\sum_{i=1}^{n} y_i x_{ik} g\left(-y_i z_i\right) + C w_k$
- $\frac{\partial^2 L}{\partial w_k \partial w_k} = -\sum_{i=1}^{n} x_{ik}^2 g\left(-y_i z_i\right) + C$
- Regularization factor C included to reduce over fitting and large parameter estimation
- Regression coefficients signifies each feature's importance in classification

Supervised Distribution Difference

❖ Overview

Introduction and Motivation

Preliminaries

❖ Notations

❖ Logistic Regression

❖ Logistic Regression Cont'd

❖ Supervised Distribution Difference

Proposed Framework

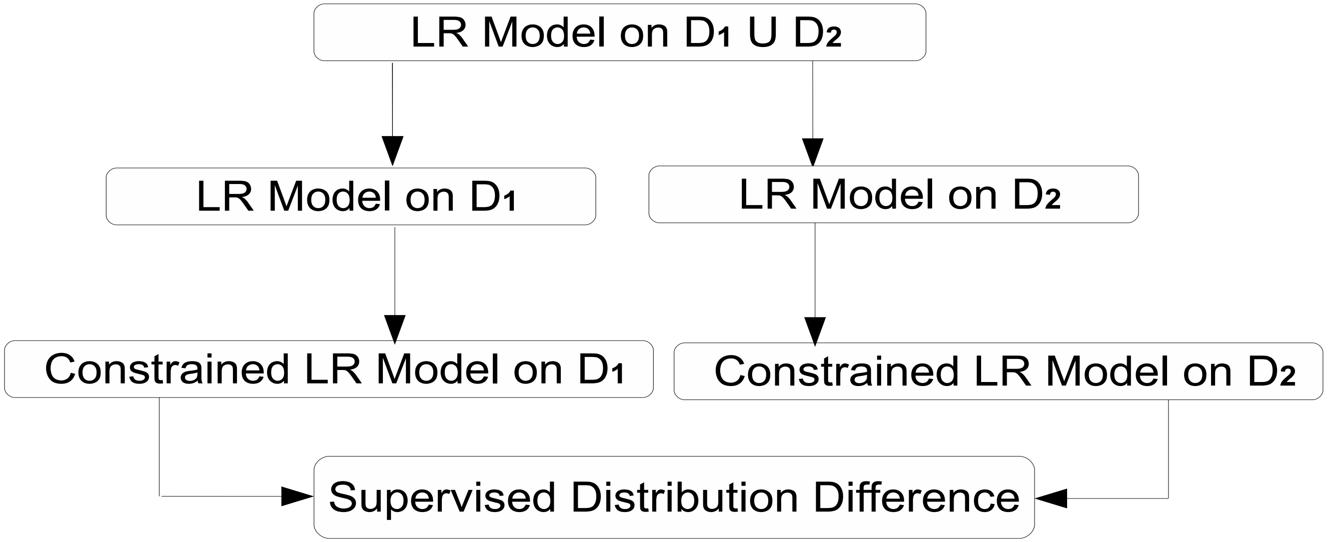
Results

Conclusion

- Supervised Distribution Difference (SDD) is defined as the change in the classification criteria in terms of measuring the deviation in classification boundary while classifying as accurately as possible.
- $SDD(\vec{w}^A, \vec{w}^B) = \sqrt{\sum_k (w_k^A - w_k^B)^2}$

Overall Approach

❖ Overview
Introduction and Motivation
Preliminaries
Proposed Framework
❖ Overall Approach
❖ Constrained Optimization
❖ Constrained Logistic Regression
❖ LR Vs Constrained LR
Results
Conclusion



- The regularization factor C for combined dataset D is obtained using 10-fold cross validation (CV)
- The complete model R on D is obtained using best regularization factor C
- Similarly LR model for D_1 and D_2 are obtained

Constrained Optimization

❖ Overview

Introduction and Motivation

Preliminaries

Proposed Framework

❖ Overall Approach

❖ Constrained Optimization

❖ Constrained Logistic Regression

❖ LR Vs Constrained LR

Results

Conclusion

- Enforce constraints on LR by restricting weight vectors

$argmin\ L = -\sum_{i=1}^{n} \log g\left(y_i z_i\right) + \frac{C}{2}\sum_{k=1}^{l} w_k^2$ subject to constraints $|R_k - w_k| \leq \ \epsilon$

- A scaled modified Newton step replaces the unconstrained Newton step [3]
- $(Z(w))^{-2}\frac{\partial L}{\partial w} = 0$
- A solution to the linear system is used to obtain solution of modified Newton step
- ϵ is the deviation we allow from individual components of weight vector
- We satisfy above equation using a constrained optimization approach on LR model

Constrained Logistic Regression

❖ Overview

Introduction and Motivation

Preliminaries

Proposed Framework

❖ Overall Approach

❖ Constrained Optimization

❖ Constrained Logistic Regression

❖ LR Vs Constrained LR

Results

Conclusion

- Calculate lower,upper bounds using ϵ
- Obtain weight vector using constrained optimization
- Model found is within τ accuracy(set to 0.15) of LR model
- If model not found, gradually increase ϵ and repeat above process until suitable model is found
- For smooth transition of models, ϵ is varied as percentage of R weight vector i.e., $\epsilon \leftarrow a \times R$

LR Vs Constrained LR

❖ Overview

Introduction and Motivation

Preliminaries

Proposed Framework

❖ Overall Approach

❖ Constrained Optimization

❖ Constrained Logistic Regression

❖ LR Vs Constrained LR

Results

Conclusion

- Constrained LR core piece is constrained minimization with box constraints
- LR essentially performs an unconstrained optimization
- The convergence proof for the termination of constrained optimization is similar to the one given in [3].

Synthetic Datasets I (SD I)

- Overview

Introduction and Motivation

Preliminaries

Proposed Framework

Results
- Synthetic Datasets I (SD I)
- Synthetic Datasets II (SD II)
- Real World Datasets (RWD)
- Validation on SD I
- Validation on SD I Cont'd
- Distribution Difference Comparison
- Sensitivity of Distance Metric
- Synthetic Datasets II (Sensitivity)
- RWD Sensitivity

Conclusion

- Two datasets generated using Gaussian distribution with predefined (μ, σ)
- Number of attributes are kept 10 in both the datasets
- Maximum class separating features are kept different in each dataset.
  - These **differential features** identify the major components responsible for difference in classification criteria
- Rest of the attributes in both the dataset are generated with similar (μ, σ)

Synthetic Datasets II (SD II)

❖ Overview

Introduction and Motivation

Preliminaries

Proposed Framework

Results

❖ Synthetic Datasets I (SD I)

❖ Synthetic Datasets II (SD II)

❖ Real World Datasets (RWD)

❖ Validation on SD I

❖ Validation on SD I Cont'd

❖ Distribution Difference Comparison

❖ Sensitivity of Distance Metric

❖ Synthetic Datasets II (Sensitivity)

❖ RWD Sensitivity

Conclusion

- A data oriented technique to generate datasets obtained by different processes introduced in [5]
- Two datasets differing purely based on data characteristics might differ in class distribution (as in this case)
- $NM.F_{num}$ denote a dataset with N million tuples generated by classification function num
- $D = 1M.F1$, $D_1 = D \cup 0.05M.F_4$, $D_2 = 0.5M.F_1, D_3 = 1M.F_2,$ and $D_4 = 1M.F_4$

Real World Datasets (RWD)

❖ Overview

Introduction and Motivation

Preliminaries

Proposed Framework

Results

❖ Synthetic Datasets I (SD I)

❖ Synthetic Datasets II (SD II)

❖ Real World Datasets (RWD)

❖ Validation on SD I

❖ Validation on SD I Cont'd

❖ Distribution Difference Comparison

❖ Sensitivity of Distance Metric

❖ Synthetic Datasets II (Sensitivity)

❖ RWD Sensitivity

Conclusion

- Five UCI datasets [1] were used in the experiments
- The binary datasets are represented by triplet (dataset, attributes, instances)
- Datasets are (blood, 5, 748), (liver, 6, 345), (diabetes, 8, 768), (gamma, 11, 19020), and (heart, 22, 267)

Validation on SD I

❖ Overview

Introduction and Motivation

Preliminaries

Proposed Framework

Results

❖ Synthetic Datasets I (SD I)

❖ Synthetic Datasets II (SD II)

❖ Real World Datasets (RWD)

❖ Validation on SD I

❖ Validation on SD I Cont'd

❖ Distribution Difference Comparison

❖ Sensitivity of Distance Metric

❖ Synthetic Datasets II (Sensitivity)

❖ RWD Sensitivity

Conclusion

Feature	*LR*	*Constrained LR*
1	-3.3732	**-0.8015**
2	-0.8693	0
3	-1.2061	-0.0158
4	-1.6274	0
5	5.0797	**0.9244**
6	1.2014	**0.4258**
7	0.0641	0.0306
8	-0.5393	0.1123
9	-3.5901	0
10	0.7765	0.0455

- Difference in individual weight vectors for two datasets for both LR and Constrained LR
- Bold features are top 3 differential features in order (1,5 and 6)

Validation on SD I Cont'd

❖ Overview

Introduction and Motivation

Preliminaries

Proposed Framework

Results

❖ Synthetic Datasets I (SD I)

❖ Synthetic Datasets II (SD II)

❖ Real World Datasets (RWD)

❖ Validation on SD I

❖ Validation on SD I Cont'd

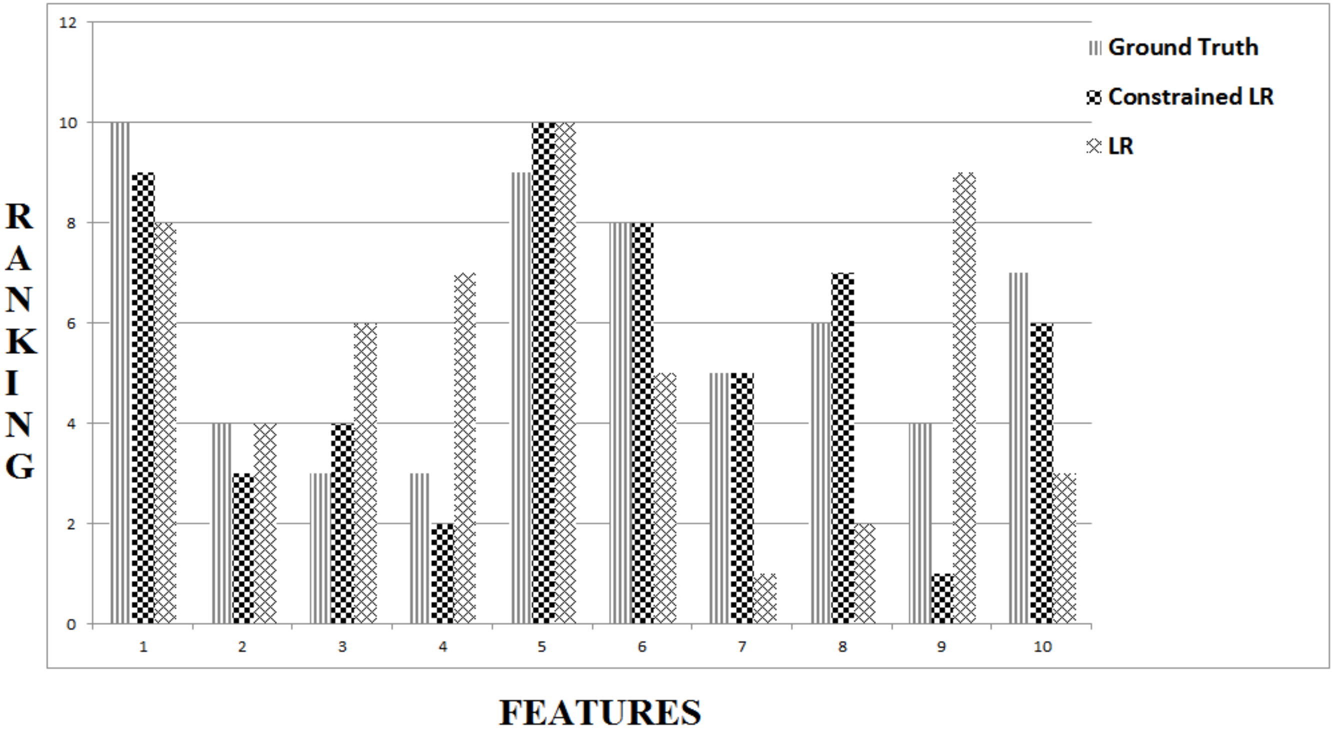
❖ Distribution Difference Comparison

❖ Sensitivity of Distance Metric

❖ Synthetic Datasets II (Sensitivity)

❖ RWD Sensitivity

Conclusion



- Constrained LR able to distinguish most differential features in correct order
- LR only able to identify two highly differential features but noisy features distort ranking

Distribution Difference Comparison

❖ Overview
Introduction and Motivation
Preliminaries
Proposed Framework
Results
❖ Synthetic Datasets I (SD I)
❖ Synthetic Datasets II (SD II)
❖ Real World Datasets (RWD)
❖ Validation on SD I
❖ Validation on SD I Cont'd
❖ Distribution Difference Comparison
❖ Sensitivity of Distance Metric
❖ Synthetic Datasets II (Sensitivity)
❖ RWD Sensitivity
Conclusion

Table 1: **The distances of all four datasets by constrained LR and Ganti's method** [5]

Dataset	*Ranking*	*Ganti's Method [5]*	*SDD*
D_1	2	0.0689	0.00579
D_2	1	0.0022	0.004408
D_3	3	1.2068	0.022201
D_4	4	1.4819	0.070124

- Relative ranking among datasets depicting difference between datasets is same.
- Only **ranking can be compared** and not distances
- Our method is able to distinguish datasets with varying degree of dissimilarity

Sensitivity of Distance Metric

❖ Overview

Introduction and Motivation

Preliminaries

Proposed Framework

Results

❖ Synthetic Datasets I (SD I)

❖ Synthetic Datasets II (SD II)

❖ Real World Datasets (RWD)

❖ Validation on SD I

❖ Validation on SD I Cont'd

❖ Distribution Difference Comparison

❖ Sensitivity of Distance Metric

❖ Synthetic Datasets II (Sensitivity)

❖ RWD Sensitivity

Conclusion

- Another way to capture differing data distribution [8]
- Create random subsamples of D of the size p
- p is varied as 10%, 20%, ..., 100%, with a stepsize of 10%
- For real world datasets, stratified sampling is suggested wherever class imbalance exists
- We expect the calculated distance between D and D_p to decrease as p increases

Synthetic Datasets II (Sensitivity)

❖ Overview

Introduction and Motivation

Preliminaries

Proposed Framework

Results

❖ Synthetic Datasets I (SD I)

❖ Synthetic Datasets II (SD II)

❖ Real World Datasets (RWD)

❖ Validation on SD I

❖ Validation on SD I Cont'd

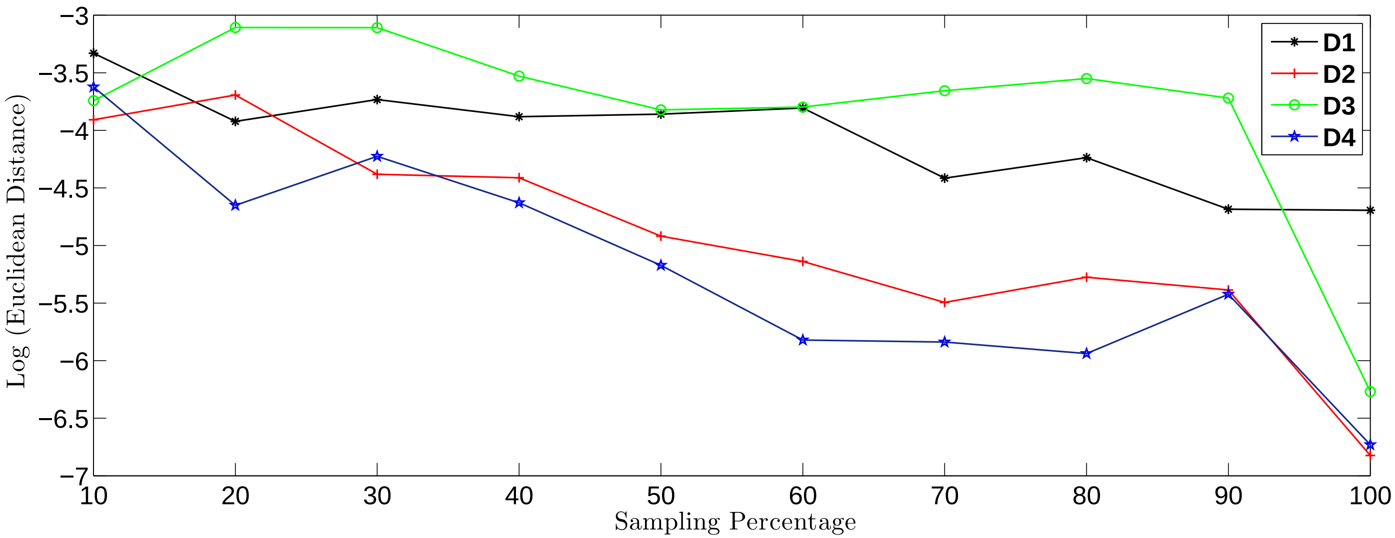
❖ Distribution Difference Comparison

❖ Sensitivity of Distance Metric

❖ Synthetic Datasets II (Sensitivity)

❖ RWD Sensitivity

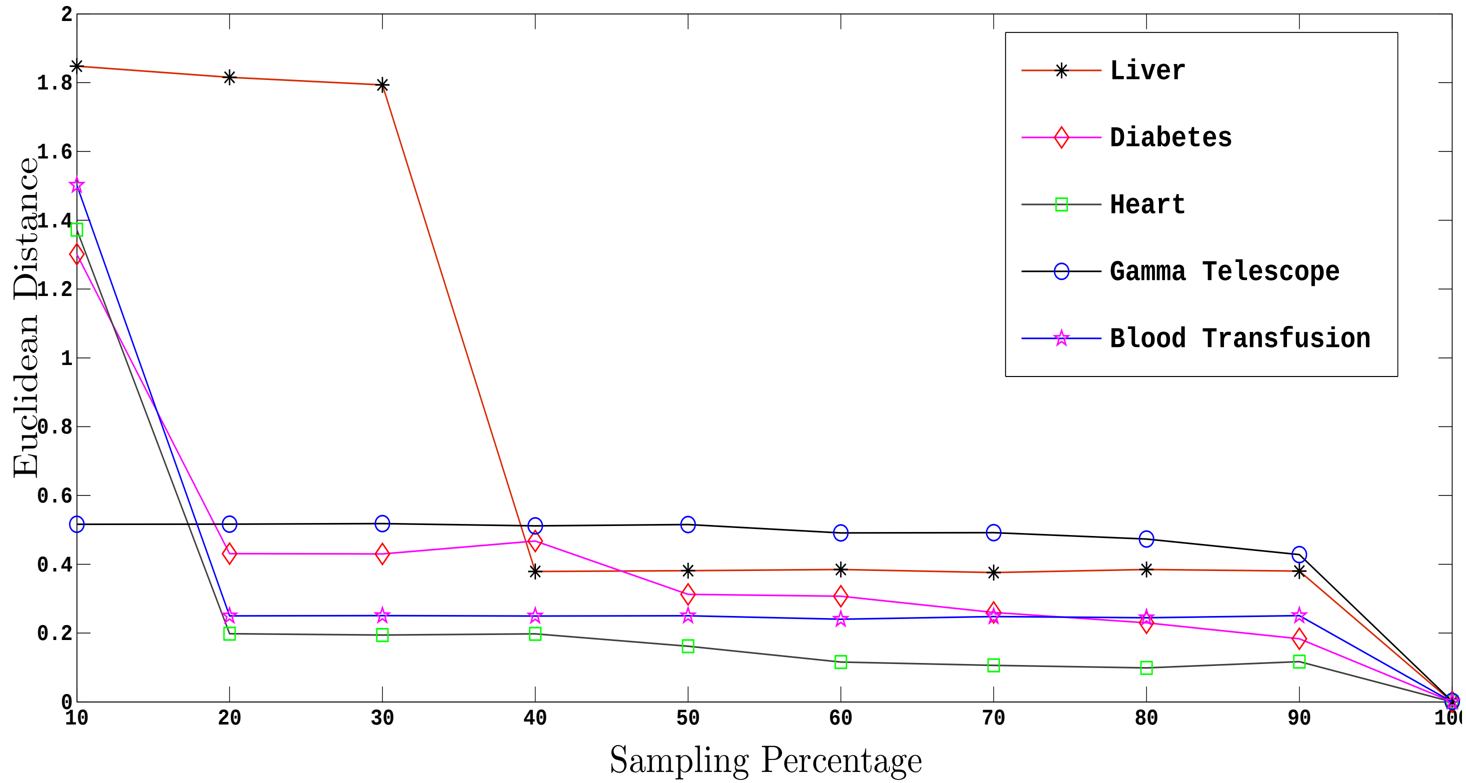
Conclusion



- Synthetic datasets are large and we observe a significant change in the class distribution even at small sampling levels
- The distance is still small and as expected decreases monotonically

Real World Datasets Sensitivity

- Overview
- Introduction and Motivation
- Preliminaries
- Proposed Framework
- Results
  - Synthetic Datasets I (SD I)
  - Synthetic Datasets II (SD II)
  - Real World Datasets (RWD)
  - Validation on SD I
  - Validation on SD I Cont'd
  - Distribution Difference Comparison
  - Sensitivity of Distance Metric
  - Synthetic Datasets II (Sensitivity)
  - RWD Sensitivity
- Conclusion



- SDD metric is significant only for 10-20% samples
- More than 20% samples in these datasets resemble class distribution of whole dataset

Conclusion and Future Works

❖ Overview

Introduction and Motivation

Preliminaries

Proposed Framework

Results

Conclusion

❖ Conclusion and Future Works

❖ References

- We developed a novel constrained logistic regression framework which captures the difference between two multivariate datasets based on the proposed distance metric.
- In this work, we considered popular linear classifier LR
- Future directions include applying kernel approaches and incorporating non-linear classifiers

References

❖ Overview

Introduction and Motivation

Preliminaries

Proposed Framework

Results

Conclusion

❖ Conclusion and Future Works

❖ References

[1] Asuncion, A., Newman, D.: UCI machine learning repository, http://archive.ics.uci.edu/ml/ (2007)

[2] Bay, S.D., Pazzani, M.J.: Detecting group differences: Mining contrast sets. Data Mining and Knowledge Discovery 5(3), 213 – 246 (2001)

[3] Coleman, T.F., Li, Y.: An interior trust region approach for nonlinear minimizations subject to bounds. Technical Report TR 93-1342 (1993)

[4] Dong, G., Li, J.: Efficient mining of emerging patterns: Discovering trends and differences. In: In Proceedings of the Fifth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining. pp. 43 – 52 (1999)

[5] Ganti, V., Gehrke, J., Ramakrishnan, R., Loh, W.: A framework for measuring differences in data characteristics. J. Comput. Syst. Sci. 64(3), 542 – 578 (2002)

[6] Lavrač, N., Kavšek, B., Flach, P., Todorovski, L.: Subgroup discovery with cn2-sd. Journal of Machine Learning Research 5, 153 – 188 (2004)

[7] Liu, B., Hsu, W., Han, H.S., Xia, Y.: Mining changes for real-life applications. In: Data Warehousing and Knowledge Discovery, Second International Conference (DaWaK) Proceedings. pp. 337 – 346 (2000)

[8] Ntoutsi, I., Kalousis, A., Theodoridis, Y.: A general framework for estimating similarity of datasets and decision trees: exploring semantic similarity of decision trees. In SIAM International Conference on Data Mining (SDM) pp. 810 – 821 (2008)

[9] Wang, K., Zhou, S., Fu, A.W.C., Yu, J.X.: Mining changes of classification by correspondence tracing. In: Proceedings of the Third SIAM International Conference on Data Mining (SDM). pp. 95 – 106 (2003)

THANK YOU

Contact Info:
rajulanand@wayne.edu
reddy@cs.wayne.edu