# Learning Complex Rare Categories with Dual Heterogeneity

**Pei Yang[1], Jingrui He[1], Jia-Yu Pan[2]**

[1]Arizona State University, {pyang33, jingrui.he}@asu

[2]Google Inc., jiayu.pan@gmail.com

# Outline

- Motivation
- Related Work
- The Proposed ***M*²*LID*** Model
- Performance Analysis
- Experiments
- Conclusion

# Motivation – Insider Threat Detection

View heterogeneity:
1) emails
2) website browsing history
3) social network

Rarity

Task heterogeneity:
the data collected from multiple financial institutes.

# Problems and Challenges

- **Rarity**

  How to effectively detect and characterize the rare categories?

- **Dual heterogeneity**

  How to leverage both task and view heterogeneity to maximally boost the performance of rare category analysis?

# Contributions

- An effective metric for boundary characterization of rare categories.
- A novel optimization framework M2LID for modeling the both rarity and dual heterogeneity.
- Performance analysis with respect to the convergence property, the error bound, and the algorithm complexity.
- Experimental results demonstrating the effectiveness of the proposed algorithm.

# Outline

- Motivation
- Related Work
- The Proposed ***M*²*LID*** Model
- Performance Analysis
- Experiments
- Conclusion

# Related Work - Rarity

- Imbalanced Classification:
  - Oversampling (Chawla et al., 2002)
  - Undersampling (Tomek, 1976)
  - One-class SVMs (Schölkopf et al., 2001)
  - Feature selection (Mladenic & Grobelnik, 1999)
  - Ensemble based methods (Zhou & Liu, 2006)

- Imbalanced Classification workshop:
  - AAAI'2000 workshop on Learning from Imbalanced Data Sets
  - ICML'2003 workshop on Learning from Imbalanced Data Sets
  - SIGKDD Explorations 2008 special issue on Learning from Imbalanced Data Sets

# Related Work - Rarity

- Outlier Detection:
  - Survey (Chandola et al., 2009)
  - Classification based (Barbara et al., 2001)
  - Nearest neighbor based (Ramaswamy et al., 2000)
  - Clustering based (Yu et al., 2002)
  - Information-theoretic methods (He et al., 2005)
  - Spectral based (Dutta et al., 2007)
  - Statistical based (Aggarwal & Yu, 2001)

# Related Work - Rarity

- Rare Category Analysis :
  - Local-density-differential sampling (He & Carbonell, 2007)
  - Active learning based sampling (Dasgupta & Hsu, 2008)
  - Hierarchical mean shift (Vatturi & Wong, 2009)
  - Gaussian mixture model (Pelleg & Moore, 2004)
  - Explore the compactness of minority with hyperball (He et al., 2010)

# Related Work – Heterogeneous Learning

- Multi-view Learning:
  - Co-training (Blum & Mitchell, 1998),
  - SVM-2K (Farquhar et al., 2005)
  - Information-theoretic method (Sridharan & Kakade, 2008)
  - Co-regularization (Sindhwani & Rosenberg, 2008)

- Multi-task Learning:
  - Feature learning based (Argyriou et al., 2007)
  - Clustered-based (Zhou et al., 2011)
  - Alternating structure optimization (Ando & Zhang, 2005)
  - Detect outlier task (Gong et al., 2012)

# Related Work – Heterogeneous Learning

- Dual (task/view) Heterogeneity:
  - Graph-based transductive method (He & Lawrence, 2011)
  - Co-regularization inductive method (Zhang & Huang, 2012)
  - Common structure learning (Jin et al., 2013)
  - Nonparametric bayes model (Yang & He, 2014)

# Outline

- Motivation
- Related Work
- The Proposed ***M*²*LID*** Model
- Performance Analysis
- Experiments
- Conclusion

# *M²LID* Model – Main Idea

- Introduce a boundary characterization metric to capture the sharp changes in density near the boundary of the rare categories in the feature space.

- Construct a graph-based model to leverage both task and view heterogeneity:
  - task-specific learners behave similarly on the features
  - view-based learners behave similarly on the examples

- M2LID models both rarity and dual heterogeneity in way of mutual benefit.

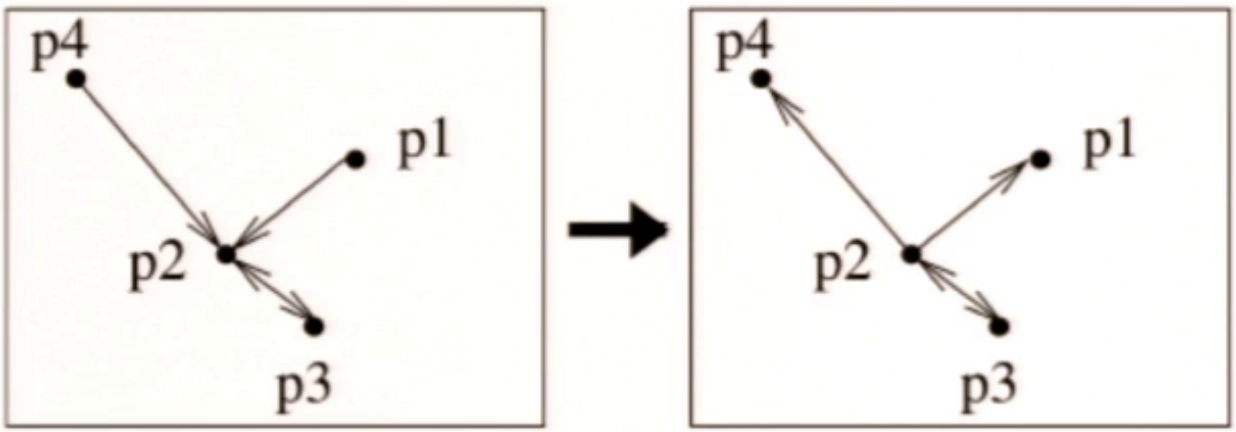

STAR Lab

# $M^2LID$ – Boundary Characterization

- Reverse K Nearest Neighbor (RKNN) vs. KNN

  The reverse k nearest neighbors of a given point is defined as (Xia et al., 2006):

$$RKNN(p_i) = \{p_j \mid p_i \in KNN(p_j)\}$$




(a) (b)

# $M^2LID$ – Boundary Characterization

- The nearest neighbor relationship is asymmetric:
- Use the different properties between KNN and RKNN to capture the sharp changes in density near the boundary of minority classes.
- If two instances have more common k-nearest neighbors, they will have more similar Hub values.
- If two instances have more common reverse k-nearest neighbors, they will have more similar Authority values.

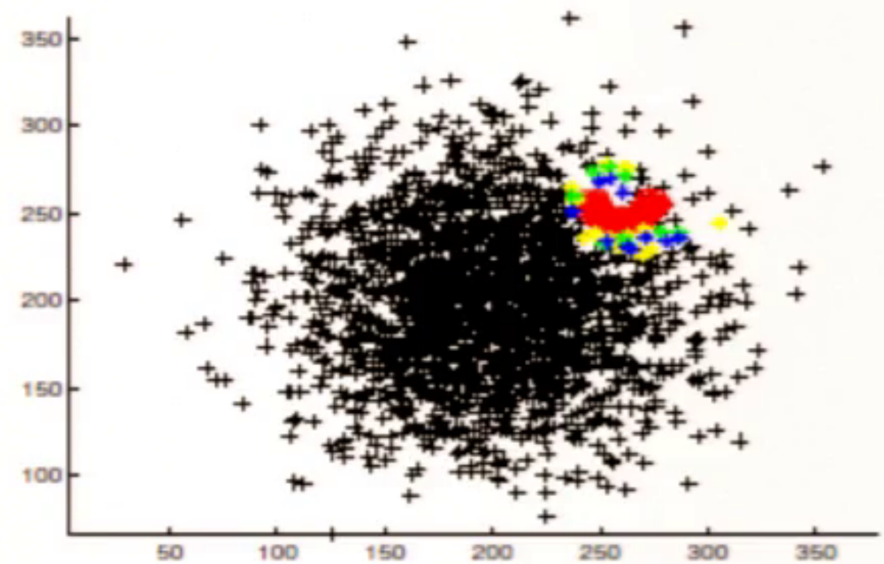
# $M^2LID$ – Boundary Characterization

- Border-degree

  Given an instance x, its border-degree is defined as:

$$b(x) = h(x) - \sigma a(x)$$

  - The larger border-degree value an instance has, the more probably it is near the boundary.
  - It is skewed around the border while flat in the regions far from border.






STAR Lab

# $M^2LID$ - Objective

- Consistency on undirected KNN graphs - Prediction:
  - smooth consistency among nearest neighbors
  - consistency with the label information
  - view consistency in terms of instances
  - task consistency in terms of features

$$J_C(f) = \sum_{i=1}^{T}\sum_{j=1}^{V} f_{ij}^T L_{f_{ij}} f_{ij} + \gamma \sum_{i=1}^{T}\sum_{j=1}^{V} \left\| f_{ij} - y_{ij} \right\|^2$$

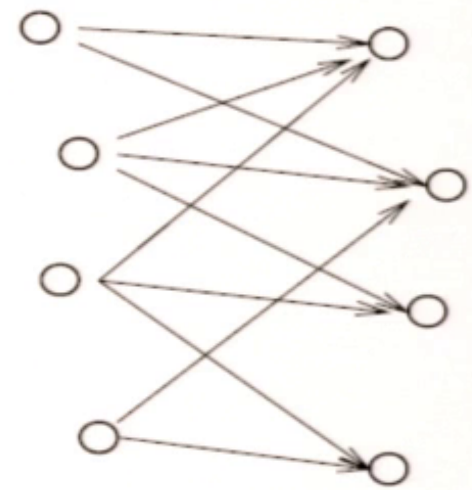
$$+ \alpha \sum_{i=1}^{T}\sum_{j,k=1}^{V} \left\| f_{ij}^I - f_{ik}^I \right\|^2 + \beta \sum_{i=1}^{V}\sum_{j,k=1}^{T} \left\| f_{ji}^F - f_{ki}^F \right\|^2$$

  - Laplace matrix $L_{f_{ij}} = L(S) = D^{-\frac{1}{2}}(D-S)D^{-\frac{1}{2}}$

# $M^2LID$ - Objective

- Hub (Kleinberg, 1999)

$$h^{t+1} = WW^T h^t$$




- Consistency on directed KNN/RKNN graphs – Hub

$$J_C(h) = \sum_{i=1}^{T}\sum_{j=1}^{V} h_{ij}^T L_{h_{ij}} h_{ij} + \alpha \sum_{i=1}^{T}\sum_{j,k=1}^{V} \left\| h_{ij}^I - h_{ik}^I \right\|^2 + \beta \sum_{i=1}^{V}\sum_{j,k=1}^{T} \left\| h_{ji}^F - h_{ki}^F \right\|^2$$

  – Laplace matrix $L_{h_{ij}} = L\left(W_{ij} W_{ij}^T\right)$

# $M^2LID$ - Objective

- Authority (Kleinberg, 1999)

$$a^{t+1} = W^T W a^t$$


- Consistency on directed KNN/RKNN graphs – Authority

$$J_C(a) = \sum_{i=1}^{T}\sum_{j=1}^{V} a_{ij}^T L_{a_{ij}} a_{ij} + \alpha \sum_{i=1}^{T}\sum_{j,k=1}^{V} \left\| a_{ij}^I - a_{ik}^I \right\|^2 + \beta \sum_{i=1}^{V}\sum_{j,k=1}^{T} \left\| a_{ji}^F - a_{ki}^F \right\|^2$$

  – Laplace matrix $L_{a_{ij}} = L\left(W_{ij}^T W_{ij}\right)$

# $M^2LID$ - Objective

- Consistency between prediction and border-degree
  - Assume y=1 for minority, y=-1 for majority;
  - Negative correlation:
    - The boundary instance have large border-degree and small absolute value of prediction.
    - The instance far away from boundary have small border-degree and large absolute value of prediction.

$$J_P(f,b) = \left[\left(\frac{f-\mu_f}{\sigma_f}\right)^2\right]^T \left(\frac{b-\mu_b}{\sigma_b}\right)^2$$

# $M^2LID$ - Objective

- Overall objective
  - Maximize the smoothness consistency objective for all of predictions, Hub, and Authority.
  - Maximize the negative correlation between the prediction and the border-degree.

$$J(f,h,a) = J_C(f) + J_C(h) + J_C(a) + \lambda J_P(f,b)$$

# The $M^2LID$ Framework

- Decision function
  - The smaller the border-degree is, the more confident the view-based classifier with its prediction.
  - The final prediction takes the weighted sum of the predictions resulting from the view-based classifiers.

$$f_i^*(x) = \sum_{j=1}^{V} \left[ 1 - \frac{b_{ij}(x)}{\sum_{k=1}^{V} b_{ik}(x)} \right] f_{ij}(x)$$


Arizona State University

# Outline

- Motivation
- Related Work
- The Proposed ***M²LID*** Model
- Performance Analysis
- Experiments
- Conclusion

# Performance Analysis

- **Convergence**

  The proposed M2LID algorithm converges to the local optimum.

$$J(f,h,a) = J_C(f) + J_C(h) + J_C(a) + \lambda J_P(f,b)$$

  - Use block coordinate descent method to optimize.
  - The objective is convex to each block {f, b, a}, e.g.,

$$J_C(f) = f^T H_f f - 2p^T f$$

$H_f$ is positive semi-definite

# Performance Analysis

$P(y=1)=r$
$P(f_j=-1 \mid y=1)=p_j$
$P(f_j=1 \mid y=-1)=q_j$

- **False Negative Error bound**

  Given the error bound,

$$\rho \geq \frac{rE\left[p_j\left(1-\bar{b}_j\right)\right]}{rE\left[p_j\left(1-\bar{b}_j\right)\right]+(1-r)E\left[\left(1-\bar{b}_j\right)\left(1-q_j\right)\right]}$$

  the probability of making a false negative error by M2LID can be bounded as follows,

$$P\left\{P\left[y=1 \mid f=-1\right] \geq \rho\right\} \leq \exp\left(\frac{-2V\mu^2}{C}\right)$$

  where

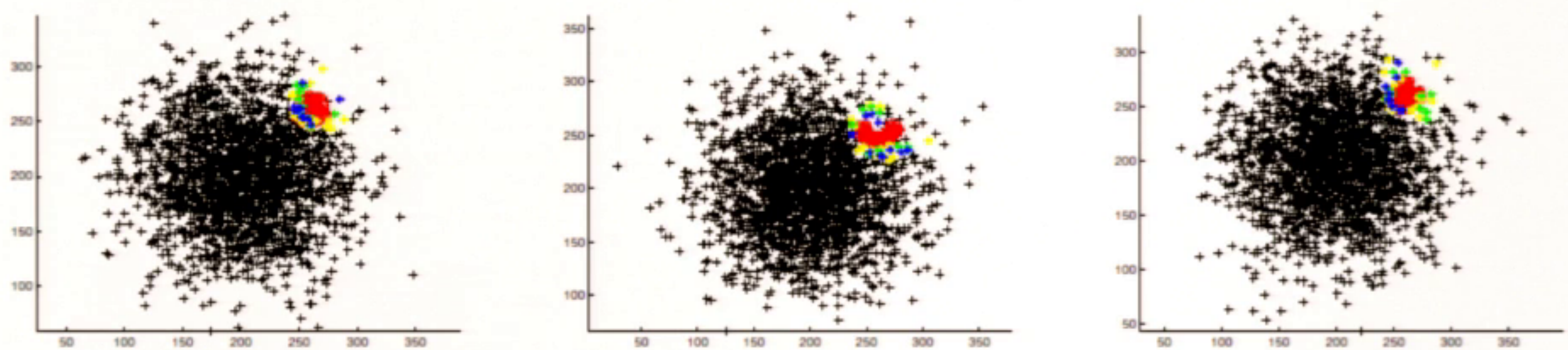
$$\mu=E\left[\left(1-\bar{b}_j\right)\left(rp_j(1-\rho)-\rho\left(1-q_j\right)(1-r)\right)\right]$$

# Outline

- Motivation
- Related Work
- The Proposed ***M*$^2$*LID*** Model
- Performance Analysis
- Experiments
- Conclusion

# Experimental Results – Synthetic Datasets

- Visualize the boundary characterization in order to verify the effectiveness of the border-degree metric:
  - 2000 majority instances ~ Gaussian distribution.
  - 100 minority instances ~ uniform distribution.
  - Three 2-dimensional datasets: Circle, Half-moon, Plus.
  - The blue (green, yellow) stars representing the instances with top-10 (20, 40) largest border-degree values.



# Experimental Results – Real Datasets

- ECML-PKDD 2006 Spam Email data
  - 3 different users (task)
  - 2500 emails per user
  - Views: TF-IDF features, topics obtained by PLSA
- Cora dataset
  - 37000 computer science research papers
  - Task refers to classify the papers in different subcategories
  - Views: TF-IDF features, topics obtained by PLSA

- Evaluation metric
  - F1-score on the minority

# Comparison with Heterogeneous Learning

- **Comparison methods**
  - Multi-task multi-view method IteM2 (He & Lawrence, 2011)
  - Multi-view method CoEM which is a variant of Co-training (Blum & Mitchell, 1998)
  - Multi-task method CASO (Chen et al., 2009)
  - Multi-task method CMTL (Zhou et al., 2011)
  - Multi-task method rMTFL (Gong et al., 2012)
  - Multi-task method RMTL (Chen et al., 2011)

# Comparison with Heterogeneous Learning

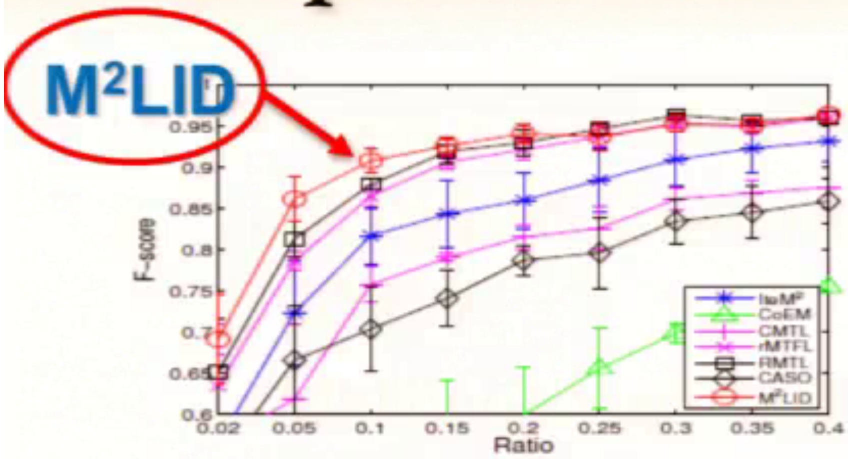



Figure 5: Error bar of different heterogeneous learning methods on Spam Email (average).

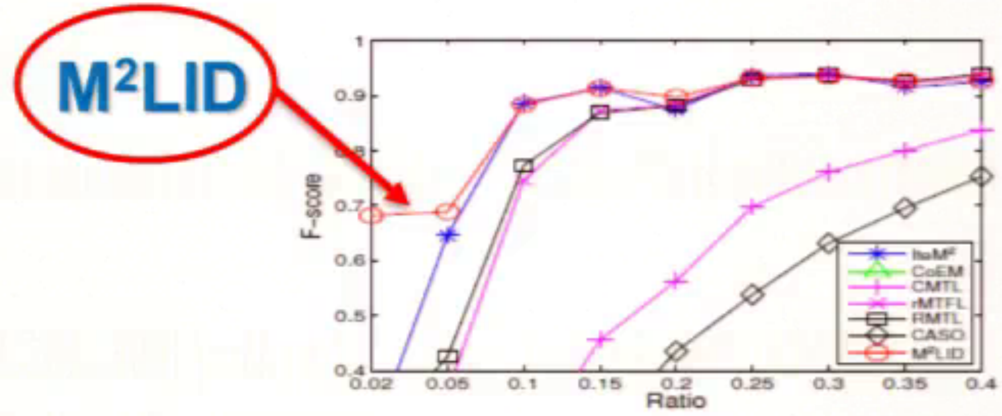



Figure 6: F-score of different heterogeneous learning methods on Cora DA-NT (average).

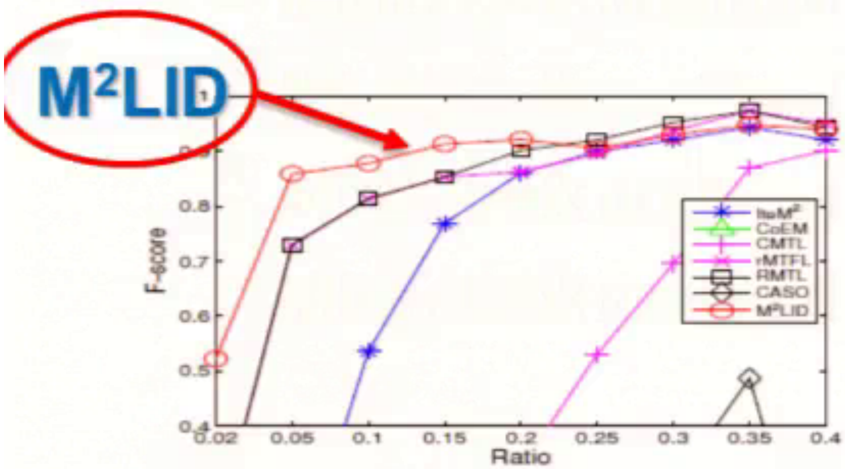



Figure 7: F-score of different heterogeneous learning methods on Cora NT-ML (average).

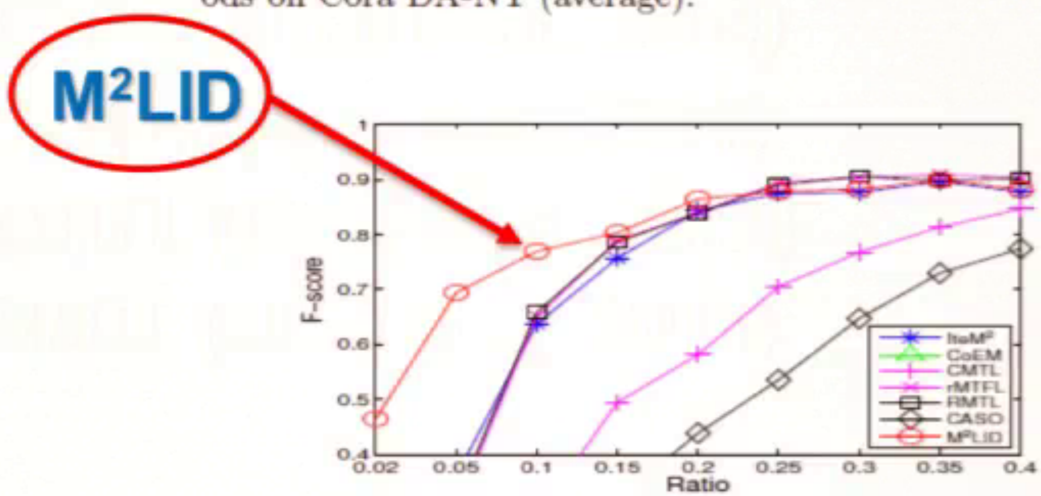



Figure 8: F-score of different heterogeneous learning methods on Cora DA-ML (average).

# Comparison with Imbalanced Learning

- **Comparison methods**
  - Oversampling
  - Undersampling
  - SMOTE (Chawla et al., 2002)
  - Ensemble methods for imbalanced data, including HardEnsemble and SoftEnsemble (Zhou & Liu, 2006).
  - All implemented in online package CSNN (http://lamda.nju.edu.cn/Data.ashx).

# Comparison with Imbalanced Learning

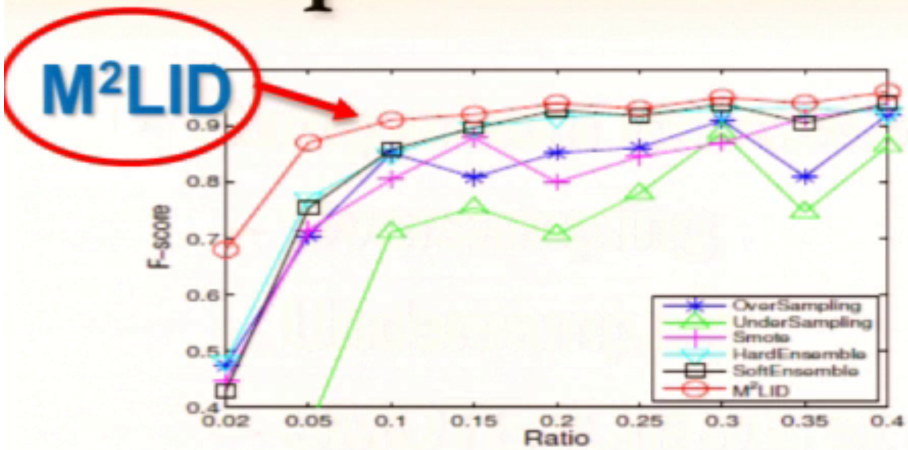



Figure 9: F-score of different imbalanced learning methods on Spam Email (average).

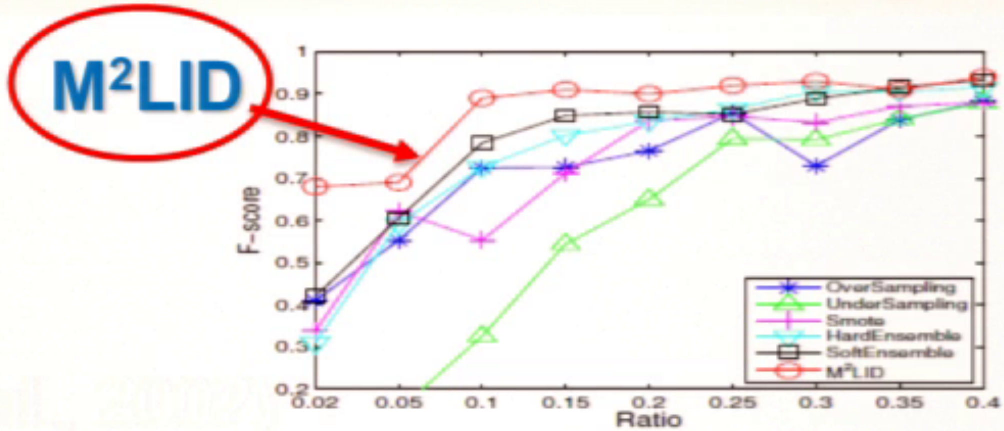



Figure 10: F-score of different imbalanced learning methods on Cora DA-NT (average).

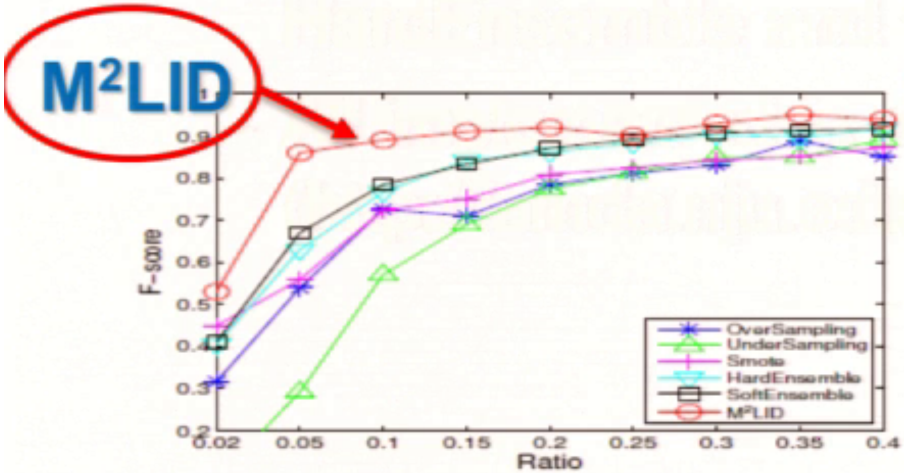



Figure 11: F-score of different imbalanced learning methods on Cora NT-ML (average).

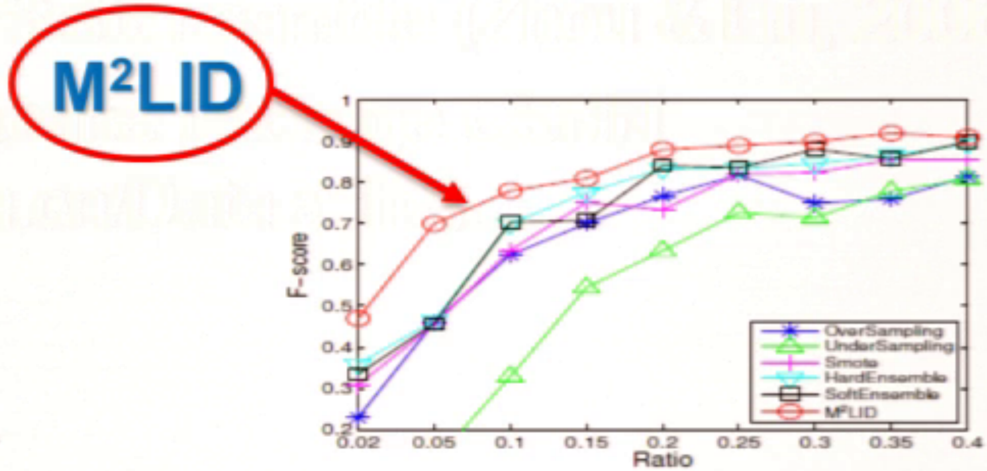



Figure 12: F-score of different imbalanced learning methods on Cora DA-ML (average).

# Parameter Sensitivity

- K is the number of nearest neighbors.
- K = 20, 30, 40, 50, 60, 70, 80, 90.
- M2LID is robust over a wide range of k values.

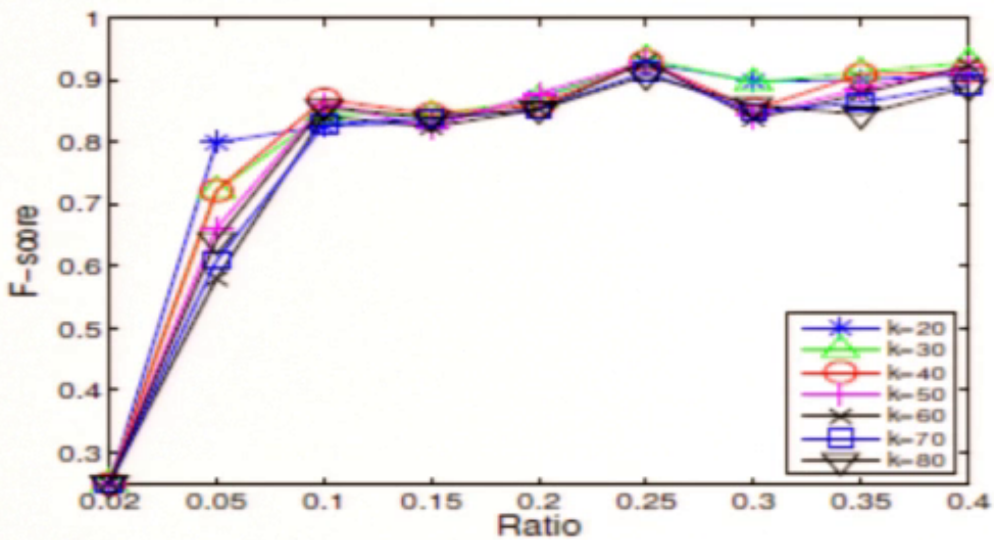



Figure 13: F-score varies with k.

# Convergence

- M2LID converges fast, and become stable after 5 iterations.

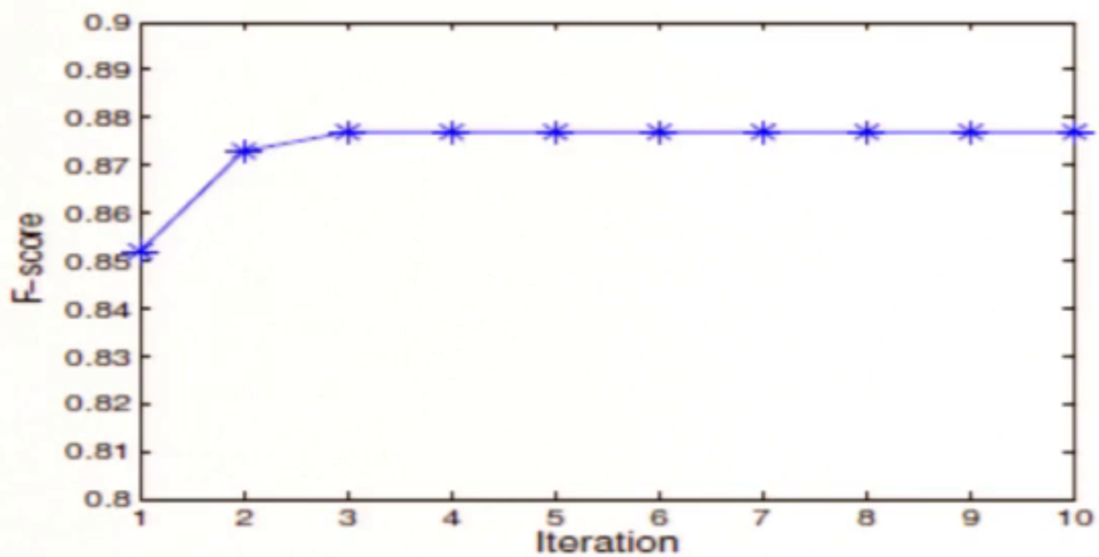



Figure 14: F-score varies with iteration.

# Outline

- Motivation
- Related Work
- The Proposed ***M*²*LID*** Model
- Performance Analysis
- Experiments
- Conclusion

# Conclusions

- An effective metric named Border-degree for boundary characterization.
- A novel M2LID framework to learn from both rarity and heterogeneity in a way of mutual benefit.
- Algorithm analysis regarding convergence, error bound, and algorithm complexity of M2LID.
- Comparisons with both heterogeneity learning and imbalanced learning methods demonstrate the effectiveness of M2LID.

Thanks