Multiple Instances Learning - EMDD

Tejas Khairnar, Vivin Paliaith {tkhairna, vivin}@asu.edu

**Abstract**—In this midterm report for CSE 569 Fundamentals of Statistical Learning we are trying to perform Image cateogarization using Estimation Maximization Diverse Denstiy Algorithm(EMDD). Multiple-Instances Learning (MIL) is a generalization of the supervised learning classification problem. Using EMDD algorithm we classify the instances in the given data set as positive instances or a negative instance. [Our algorithm calculates the probability of the instance to be a positive or negative instance].

**Index Terms**—Multiple-Instances, Diverse Density, Estimation Maximization.

xxxx-xxxx/0x/$xx.00 © 200x IEEE Published by the IEEE Computer Society

————————————————

1. F.A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305. E-mail: author@ boulder.nist.gov.
2. S.B. Author Jr. is with the Department of Physics, Colorado State University, Fort Collins, CO 80523. E-mail: author@colostate.edu.
3. T.C. Author is with the Electrical Engineering Department, University of Colorado, Boulder, CO 80309. On leave from the National Research Institute for Metals, Tsukuba, Japan E-mail: [author@nrim.go.jp](mailto:author@nrim.go.jp).

**\*\*\*Please provide a complete mailing address for each author, as this is the address the 10 complimentary reprints of your paper will be sent**

*Please note that all acknowledgments should be placed at the end of the paper, before the bibliography (****note that corresponding authorship is not noted in affiliation box, but in acknowledgment section****).*

—————————— ◆ ——————————

# 1 Introduction

M

ultiple Instances Learning (MIL) is proposed as a variation of supervised learning for problems with incomplete knowledge about lables of traning examples. In supervised learning, every traning instance is assigned with a discrete or real-valued label. In comparison, in MIL the labels are only assigned to bags of instances. In the binary case, a bag is labelled positive if at least one instance in that bag is positive, and the bag is labelled negative if all the instance in it are neagitve. There are no lables on the individual instances. The goal of MIL is to classify unseen bags or instances based on the labelled bags as the training data. For this specific project we focus on EMDD algorithm to perform the traning and categorization. However, we also compare this implementation of ours with various other implementations of MIL algorithms.

# 2 Understanding

## 2.1 Diverse Density algorithm

The main idea of DD approach is to find a concept point in the deature space that are close to at-least one instance from every positive bag and meanwhile far away from instances in negative bags [1]. The optimal concept point is defined as the one with maximum density, which is a measure of how many different positive bags have instances near the point, and how far the negative instances are from that point.

## 2.2 Estimation Maximization

The Esitmation Maximixation algorithm is an effective iterative procedure to compute the Maximum Likelihood (ML) estimate in the presence of missing or hidden data [2]. In ML estimation, we wish to estimate the model parameter(s) for which the observed data are the most likely.

Each iteration of the EM algorithm consists of two processes: The E-step, and the M-step. In the expectation, or E-step, the missing data are estimated given the observed data and current estimate of the model parameters. This is achieved using the conditional expectation, explaining the choice of terminology. In the M-step, the likelihood function is maximized under the assumption that the missing data are known. The estimate of the missing data from the E-step are used in lieu of the actual missing data.

Convergence is assured since the algorithm is guaranteed to increase the likelihood at each iteration.

## 2.3 EMDD

EM-DD [3] starts with an initial guess of the concept point t (which can be obtained using original DD algorithm), and then repeatedly performs the following two steps: E-step, the current hypothesis of concept t I used to pick the most likely instances from each bag given a generative model; in M-step, a new concept t’ is estimated by maximizing a transformed DD defined on the instances selected in the E-step using the gradient search. Then, the old concept t is replaced by the new concept t’ and the two steps are repeated until the algorithm converges. This EM-DD algorithm which we have implemented, implementes a “hard” version of EM, since only one instance per bag is used for estimating the hypothsis. It can be also regarded as a special case of the K-means clustering algorithm[4], where only on cluster is considered.

# 3 Progress

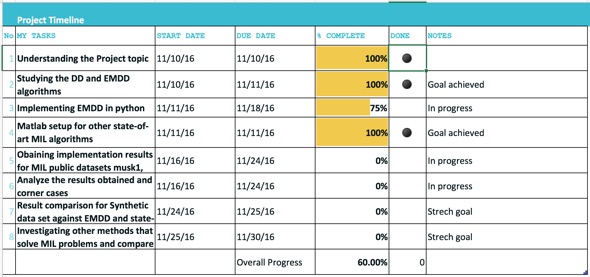
Until midterm, we have successfully implemented the EMDD algorithm in python programming language. In order to perform the Math behind the algorithm we are using the SciPy library [5]. According to the goals allocated for this project, we have tested our implementation against the Synthetic data set [6]. In order to perform the comparison of results obtained by our implmenetation and the Matlab implementation of MIL algorithms [7], we performed the Matlab setup and ran the Matalb setup for algorithms like

1. Iterated-discrim APR
2. Diverse Density
3. Two SVm variants for MIL
4. Citation-kNN for MIL

Currently we have the results for these algorithms against the Syntheic dataset. We are now in progess of moving toward the public dataset list [7] [8] [9].

# 4 Timeline

Below is the timeline for this project which we chalked out:



# 5 Results

The workload amongst us was shared equally. We worked together on the implementation. The major part of coding was done by Vivin and I took care of the Matlab setup and result comparion part.

# 6 Workload shared

The workload amongst us was shared equally. We worked together on the implementation. The major part of coding was done by Vivin and I took care of the Matlab setup and result comparion part.

# Conclusion

Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions. Authors are strongly encouraged not to reference multiple figures or tables in the conclusion—these should be referenced in the body of the paper.

**Acknowledgment**

The authors wish to thank A, B, C. This work was supported in part by a grant from XYZ.

**References**

1. J.S. Bridle, “Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition,” *Neurocomputing—Algorithms, Architectures and Applications,* F. Fogelman-Soulie and J. Herault, eds., NATO ASI Series F68, Berlin: Springer-Verlag, pp. 227-236, 1989. (Book style with paper title and editor)
2. W.-K. Chen, *Linear Networks and Systems.* Belmont, Calif.: Wadsworth, pp. 123-135, 1993. (Book style)
3. H. Poor, “A Hypertext History of Multiuser Dimensions,” *MUD History,* http://www.ccs.neu.edu/home/pb/mud-history.html. 1986. (URL link \*include year)
4. K. Elissa, “An Overview of Decision Theory,"unpublished. (Unpublished manuscript)
5. R. Nicole, "The Last Word on Decision Theory," *J. Computer Vision,* submitted for publication. (Pending publication)
6. C. J. Kaufman, Rocky Mountain Research Laboratories, Boulder, Colo., personal communication, 1992. (Personal communication)
7. D.S. Coming and O.G. Staadt, "Velocity-Aligned Discrete Oriented Polytopes for Dynamic Collision Detection," *IEEE Trans. Visualization and Computer Graphics*, vol. 14,  no. 1,  pp. 1-12,  Jan/Feb  2008, doi:10.1109/TVCG.2007.70405. (IEEE Transactions )
8. S.P. Bingulac, “On the Compatibility of Adaptive Controllers,” *Proc. Fourth Ann. Allerton Conf. Circuits and Systems Theory*, pp. 8-16, 1994. (Conference proceedings)
9. H. Goto, Y. Hasegawa, and M. Tanaka, “Efficient Scheduling Focusing on the Duality of MPL Representation,” *Proc. IEEE Symp. Computational Intelligence in Scheduling* *(SCIS ’07)*, pp. 57-64, Apr. 2007, doi:10.1109/SCIS.2007.367670. (Conference proceedings)
10. J. Williams, “Narrow-Band Analyzer,” PhD dissertation, Dept. of Electrical Eng., Harvard Univ., Cambridge, Mass., 1993. (Thesis or dissertation)
11. E.E. Reber, R.L. Michell, and C.J. Carter, “Oxygen Absorption in the Earth’s Atmosphere,” Technical Report TR-0200 (420-46)-3, Aerospace Corp., Los Angeles, Calif., Nov. 1988. (Technical report with report number)
12. L. Hubert and P. Arabie, “Comparing Partitions,” *J. Classification,* vol. 2, no. 4, pp. 193-218, Apr. 1985. (Journal or magazine citation)
13. R.J. Vidmar, “On the Use of Atmospheric Plasmas as Electromagnetic Reflectors,” *IEEE Trans. Plasma Science*, vol. 21, no. 3, pp. 876-880, available at http://www.halcyon.com/pub/journals/21ps03-vidmar, Aug. 1992. (URL for Transaction, journal, or magzine)
14. J.M.P. Martinez, R.B. Llavori, M.J.A. Cabo, and T.B. Pedersen, "Integrating Data Warehouses with Web Data: A Survey," *IEEE Trans. Knowledge and Data Eng.*, preprint, 21 Dec. 2007, doi:10.1109/TKDE.2007.190746.(PrePrint)