BEST: International Journal of Management Information Technology and Engineering (BEST: IJMITE)
ISSN (P): 2348–0513, ISSN (E): 2454–471X
Vol. 9, Issue 2, Dec 2021, 47–56
© BEST Journals

IMPORTANCE OF ACTIVE LEARNING FOR OPTIMAL DETECTION OF DISEASES

CHANDNI CHATTERJEE, PRANAB HAZRA, SUVADEEP GUHA*, AVIJIT SAHA & RAJU DAS

Department of Electronics and Communication, Narula Institute of Technology, Kolkata, India

ABSTRACT

Availability of data is very easy and most of the data are unlabeled. If we want to take some decision from the available data, it is required to be processed and data should be labeled. This unlabeled data is a big problem in the field of machine learning. Data scientists are proficient for analyzing with more data than they have and it's the point where active learning comes into picture. In machine learning, active learning acts as a subset where learning algorithm interact with user by putting query to label data with the desired outputs. In active learning a query unit is of the same type as the target concept to be learned. Alternative query is introduced in the context of multiple-instance active learning (MIL) where instances are grouped into bags, and it is the bags, rather than instances, that are labeled for training. A bag is labeled negative if and only if all of its instances are negative and positive, even if at least one of its instances is positive. In this paper we can see at higher specificity the Area Under Receiver operating characteristics (ROC) Curve (AUC) increases. Combined use of MIL and AL will reduce the labeling effort and it will do accurate detection of disease through classification.

KEYWORDS: Active Learning, Multiple Instance Learning, Machine Learning, Bag, ROC, Labeled Data, Unlabeled Data.

I. INTRODUCTION

Active learning is a subfield of machine learning. It contains queries in the form of unlabelled data which has to be labelled by the oracle (human annotator). Its main objective is to reduce the manual effort in processing the data of machine learning classifier and to acquire greater accuracy by applying fewer training labels.

Active learning has become aspiring in several trendy modern machine learning problems where data may be abundant but labels are scarce or expensive to obtain.

II. ACTIVE LEARNING SCENARIOS

Following are some problem scenarios of active learning:-

- Membership query synthesis
- Stream-based selective sampling
- Pool-based sampling.

Membership Query Synthesis

In this scenario, the unlabelled instance that are placed as input, are asked to be labelled by the learner including the queries that the learner generates de novo, rather than only those which are already sampled from some underlying natural distribution.

For example: If the data is pictures of alphabets, the learner would create an image that is similar to that picture of digits. This created image is then sent to the oracle for labeling.

Stream-Based Selective Sampling

In this setting, assumptions are made in getting an unlabelled instance free. On the basis of this assumption, unlabeled instances are selected one at a time and permit the learner to see whether it desires to query the label of the instance or reject it based on its informativeness. Query strategies are used to determine the informativeness of the instance.

For instance, an image selected from a set of unlabelled images is firstly determined as whether it needs to be labelled or discarded, and then repeat with the next image.

Pool-Based Sampling

This setting assumes that there is an outsized pool of unlabelled data. Instances are then drawn from the pool on the basis of informativeness measure. The most informative instance is then selected for querying. For example, if the data is a picture of alphabets, all the unlabelled instances from the picture of alphabets will be ranked first and then the most informative instances will be selected and are requested to be labeled by the oracle.

III. PRACTICAL CONSIDERATIONS

Batch-Mode Active Learning

In this type of active learning, learners are allowed to query instances in group rather than selecting them one at a time. The size of batching depends on the learner. This batch-mode of active learning is often preferable to sequential methods when each label takes substantial time but can be produced in parallel. The challenge in batch-mode active learning is how to properly assemble the optimal query set Q. There are two parts in this method. One is Batch Rank and other one is Batch Rand. At first, we have to formulate the batch selection as an NP hard which is integer quadratic programming problem. Then we call Batch Rank method which is linear programming. Then we also used another method that is Batch Rand which is semi definite programming. By using these two methods we can able to solve the Batch Mode Active Learning problem. So now we consider a Batch Mode Active Learning problem. At first, we take training set of data which is represented by $L_{t.}$ It is carry the information. Then we take U_t. It is represent an unlabelled data set and t is time. After that we take batch or set which is represented by B. It contains k points. Suppose in the set there are fifty thousand of data but from these data we take just fifty unlabeled data which is related to the information and k will represent these data. So the value of k is fifty. Now need classifier which is represented by the ω^t and also need entropy vector which is represented by C. Now R is real numbers. Now data will need from that unlabelled will be applied as an input data set and as subject to the model ω^t then we get an output which is represented by y.

$$c(i) = S(y|x_{i,}\omega^t) = -\sum P(y|x_i, \omega^t)\log P(y|x_i, \omega^t \qquad (1)$$

Support Vector Machine

SVM is a supervised model used for classification problems. The proposed approaches were compared against (a) Random Selection, (b) Distance based Selection, in which an SVM was trained for each possible class and the k closest unlabeled points (k is batch size) to all the hyper planes in the feature space were selected for annotations, (c) Entropy based Selection, where the entropy was computed for every unlabeled instance and the top k points were queried based on the entropy ranking. Here for divide into the two datasets need Non-linear Support Vector Machine. This Non-linear Support Vector Machine divided two datasets by using Kernal function. A polynomial Kernal Support Vector Machine was used as the basic classifier because of its established performance in the multi-label learning. The Kernal function takes as input low dimensional features space(mix data) and give the output as high dimensional feature space and the classification also done by the hyper plane figure 1.

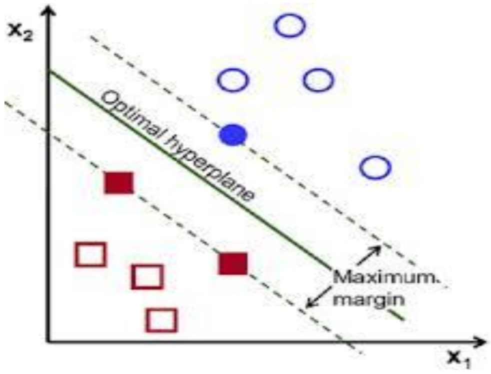

Figure 1: Support Vector Machine.

Alternative Query Types

Active learning deals with assigning class labels to text documents, the learner is required to query an unlabeled data and the oracle provides its label. Alternative query is introduced in the context of multiple-instance active learning. In multiple-instance (MI) learning, instances are grouped into bags (i.e., multi-sets), and it is the bags, rather than instances, that are labelled for training. A bag is labelled negative if and only if all of its instances are negative. A bag is labelled positive, however, if at least one of its instances is positive.

Multi-Task Active Learning

We can do lot of things using machine learning models in general such as image classification, object detection, text generation and so on. Typically a model is trained to do a single task, which is convenient, but we may want to use a single model to solve multiple problems for various reasons, such as efficiency and better generalization. To improve efficiency, one thing we can do is to share some of the layers between different but related task.

Therefore, a single query will be labelled for multiple tasks, and attempt to assess the informativeness of a query with respect to all the learners involved.

IV. STUDY ON THE DETECTION OF TUBERCULOSIS USING MIL, ACTIVE LEARNING AND ONE CLASS CLASSIFIER

Detection of Tuberculosis can be done by applying a widely utilized tool called chest radiography which is generally used during the early detection of TB. Chest radiography (CRX) involves performing X-ray examination that detects lungs abnormalities like TB, pneumonia, lungs cancer etc. Though chest radiography has high sensitivity[15] for pulmonary TB, it is difficult to achieve the performance level as it requires the influence of qualified and knowledgeable personnel, which is lacking in many resource constraint countries. To overcome this drawback computer aided detection (CAD) has been introduced as this technology enables automatic assessment of CRX following multiple instance learning (MIL) approach. Comparatively CAD performs better than clinical officers. The major advantages of CAD includes their low cost, easy operation, better quality, improves accuracy and many more. Even in underdeveloped areas, modern digital radiography machines are very affordable. Traditionally CAD was performed manually by outlining the lesions and applying training and optimization of large database. Then again this approach requires a lot of effort and also time consuming. In such case MIL can be used. It is a supervised labeled bags. A bag is a collection of instances. It is labeled negative, if all the instances comprising in it are negative and labeled positive even if at least one of the instance is positive.

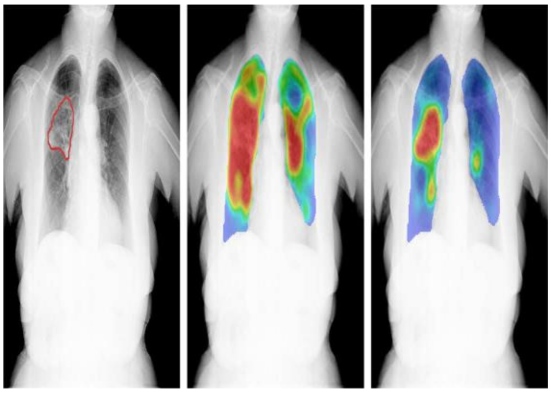

Figure 2: Improvement in Detection of TB from the Image Formed by CRX (left), MIL Approach (Middle) and Further Supervision by MIL Approach [16].

Unfortunately, MIL is not a completely conflict free technique. It is known that negative bags are entirely composed of only normal pixels, but the positive bags consist of both normal and abnormal instances which imply that although the score provided for the image is accurate the individual pixel (instance) score is not up to the mark. For instance, from Fig 2 it can be observed that the centre image has higher positive instance than the one in the left with manually outlined lesion generated by CRX. This shows that the centre image contains false-positive detections within the healthy respiratory organ. In the utmost right case, the suspicious region is highlighted properly with minor false detected positive region.

In this study, we have a tendency to propose an improved algorithm of MIL classifier which will overcome the antecedent outlined drawbacks by introducing other paradigms of machine learning like active learning (AL) and one class classification to make the output a lot more correct. This proposed algorithm aims at boosting the instance classification performance.

V. PROPOSED METHOD

A training stage has been proposed to improve the MIL classifier by combining it with active learning and one-class classifier. The steps involved are: MIL training and classification, selection of valuable instances.

A. MIL Training and Classification

The MIL classifier utilized here corresponds to the si-miSVM+PEDD technique developed in [17] which is responsible for detection of tuberculosis on chest X-rays and is dependent on the miSVM formulation defined in [18].

$$min_{\{y_i\}} min_{w,b,\xi} \frac{1}{2} ||w||^2 + C\sum_i \xi_i$$

$$\text{s.t. } y_i((w,x_i)+b) \geq 1-\xi_{i,} \xi_i > 0$$

$$\forall_i : Y_I=1, \sum_{i \epsilon I} \frac{y_i+1}{2} \geq 1,$$

$$\forall_i : Y_I = -1, y_i = -1,$$

$$y_i \text{€} \{ -1,1\} \tag{2}$$

In equation (2) y_i belongs to {-1, 1} and i = 1,...,N, are the labels of the training instances x_i ; $Y_{I,}$ Y_I € {-1,1}, are the label of the training bags $B_{I.}$ Index set I, B_I= { x_i : i**x** I } for the Index set I ⊆{1,...,N }. w is the weight vector, b is the offset of the separating hyperplane and C is the penalization parameter for the misclassified instances and also **€ᵢ** are slack variables as in the standard soft margin SVM [20].

si-miSVM + PDD in [17] is the improvised version of the one proposed in [18] used for performing the optimization in equation (1) whereas the original technique miSVM in [18] is utilized for recovering the individual instance present in the bags. First step of this proposed method is finding the confidence value by classifying the training set which later will be used to decide the target value for relabeling.

B. Selection of Valuable Instances

The main motive of active learning here is to build an explicit classifier. A small set of unlabelled instances containing the most valuable instances have been selected to be labelled by the oracle (human annotator). Two types of selecting criterion is mostly used in active learning: informativeness and representativeness [29] where Informativeness measures whether an instance is capable of reducing the uncertainty of a model or not, while representativeness measures whether the presentation of the input pattern is well represented by the instance or not. We have used here both the cases instead of just using the conventional informativeness. Instances are queried based on which MIL classifier is the most confident. Usually, in active learning, instances are labelled individually. Although that approach will not be applied here, rather labelling a group of instance will be more appropriate as analyses is done mainly on image region and not on individual instances. Thus, the approach of grouping a number of instances will be carried out in this study by utilizing mean shift clustering algorithm [19].

Even after instance grouping, there may be still some region left to be label which can be done by following two criterions.

- Selecting the region having the most instances, so that more numbers of instances get correctly labeled.
- Focusing on the regions which have greater chance to be normal so that the false-positive instances in the positive bags can be correctly relabeled as negative.

The first criterion can be easily accomplished by sorting the regions on the basis of their number of instances. Performing the second criterion will not be that easy as it deals with problem that whether the label given to the training set is correct or not. One-class classification plays a role here by assuming the information of a specific class which is also known as the target class, and obtain the information about the objects of the class.

Firstly, one-class classifier will perform the training on normal region and then it will be performed on positive region containing both normal and abnormal region. One class classifier identifies the instances as which instances are actually positive and which are not.

For performing the one-class classification k- nearest neighbour method [16] has been put to use. Here, x_j, j=1,....,M be the training instance present in the region from where abnormal image is obtained as R_j ={x_j : j∈ J} for J⊆ $\{1, \ldots, M\}$.

- Sorting the region in downward sloping order according to the number of instance.
- Obtaining the normality score of the instance with respect to the nearest neighbour as defined in equation 3.

$$\eta_j = \| x_j - x_k \| \tag{3}$$

In equation 4. instances are scaled into [0,1] interval using the minimum and maximum values observed from the normality scores:

$$\eta'_j = \frac{\eta_j - min_j \eta_j}{max_j \eta_j - min_j \eta_j} \tag{4}$$

In equation 5. Computation of normality score per region is done

$$\eta_J = \frac{1}{|J|} \sum_{j \in J} n'_j \tag{5}$$

- Identification of regions whose normality score below a certain threshold is performed and those with low normality score is then moved to the bottom of the ranking. The top regions are then selected for relabeling by an expert.

I. OBSERVATIONS

A combination of MIL and AL with 100 and 300 labelled regions is compared here with pure MIL and SVM. The performance of detecting the comparison is carried out by following the receiver operating characteristic (ROC) approach. In figure 3 the first column consist of CRX of the lungs affected by Tuberculosis and the second column is the heat map produced by applying an MIL algorithm. The column 3 of figure 3 is a heat map produced by following an SVM methodology. Column 4 and column 5 in figure 3 is produced by combining MIL and active learning (AL) keeping 100 and 300 as the ranked region respectively. From figure.3 it is observed that there have been a constant improvement in eliminating the normal region from the abnormal region.

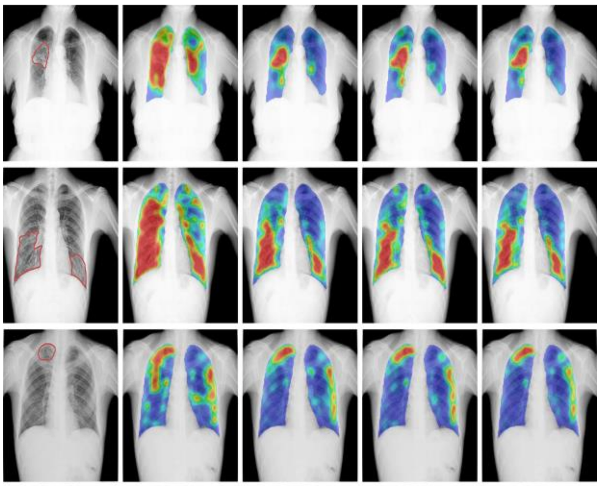

Figure 3: Image of TB Affected Lungs Obtained by CRX and Heat Map [16].

The usage of 100 and 300 labelled regions with the combination of MIL and AL have contributed even more. The reduction of false positive cases is visible in both normal and abnormal cases. From figure 4 we can notice that the Area Under ROC Curve (AUC) is larger when 100 region is included and even more when 300 is added. Thus from this we can conclude that at higher specificity the AUC increases and hence make the improvement more prominent. At lower specificity the difference in the performance of MIL, MIL+AL and SVM is barely visible. Thus, at these operating points the labelling is not very essential. Major advantage of the proposed method is the reduced labelling effort, especially when low numbers of region is utilized. Usage of one-class classification has enhanced the effectiveness of the proposed method.

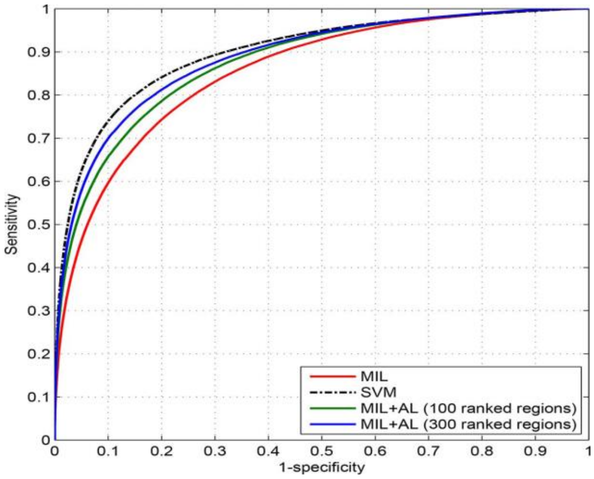

Figure 4: Gradual Improvement of TB Detection Sensitivity from MIL to MIL+AL and further Supervision by MIL+AL [16].

VII. CONCLUSIONS

The purpose of this study aims at applying active learning and its application for optimal ways of disease detection. Active learning has now become a growing space in machine learning algorithm. Practical applications of active learning such as batch mode active learning, multiple instance learning has been put into focus here. A study is proposed here by combination MIL, active learning and one class classifier to reduce the unreliability of MIL classifier and to make the work more accurate. Thus from this we can conclude that at higher specificity the AUC increases and hence make the improvement more prominent. At lower specificity the difference in the performance of MIL, MIL+AL and SVM is barely visible. Here the labelling of operating point is not very essential. From the observation, it can be concluded that this study is suitable for detection diseases like Tuberculosis.

ACKNOWLEDGEMENT

The authors acknowledge with thanks the support from Narula Institute of Technology, Agarpara, Kolkata, India for encouraging in research work and giving scope in this regard. We also thankful to the department of Electronics and Communication Engineering of Narula Institute of Technology for carrying out the study.

REFERENCES

1. S.C. Hoi, R. Jin, and M. Lyu, “Batch mode active learning with applications to text categorization and image retrieval,” IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1233–1248, Sep. 2009.
2. S. C. H. Hoi, R. Jin, J. Zhu, and M. R. Lyu, “Batch mode active learning and its application to medical image classification,” in Proc. 23rd Int. Conf. Mach. Learn., 2006, pp. 417–424.
3. S. C. H. Hoi, R. Jin, and M. R. Lyu, “Large-scale text categorization by batch mode active learning,” in Proc. 15th

Int. Conf. World Wide Web, 2006, pp. 633–642.

4. D. Zhang, F. Wang, Z. Shi, and C. Zhang, “Interactive localized content based image retrieval with multiple-instance active learning,” *Pattern Recogn.*, vol. 43, pp. 478–484, 2010.

5. C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Trans. Intell. Syst. Technol.* vol. 2, pp. 27:1–27:27, 2011.

6. A. Kapoor, E. Horvitz, and S. Basu, “Selective supervision: guiding supervised learning with decision-theoretic active learning,” in Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, Jan. 2007, pp. 877–82.

7. Y. Guo and D. Schuurmans, “Discriminative batch mode active learning,” in Proc. Adv. Neural Inf. Process. Syst., 2007, pp. 593–600.

8. Y. Guo, “Active instance sampling via matrix partition,” in Proc. Adv. Neural Inf. Process. Syst., 2010, pp. 802–810.

9. M. Kukar, “Transductive reliability estimation for medical diagnosis,” J. Artif. Intell. Med., vol. 29, pp. 81–106, 2003.

10. J. Guo, H. Chen, Z. Sun, and Y. Lin, “A novel method for protein secondary structure prediction using dual-layer SVM and profiles,” Proteins, Vol. 54, pp. 738–43, Mar. 2004. doi: 10.1002/prot.10634 [Crossref], [PubMeb],[Web of science], [Google Scholar]

11. R. Hu, “Active learning for text classification,” PhD thesis, Dublin Institute of Technology, 2011. [Google Scholar]

12. H. T. Nguyen and A. Smeulders, “Active learning using pre-clustering,” in Proc. ICML, 2004, pp. 623–630.

13. A. McCallum, and K. Nigam, “Employing em and pool-based active learning for text classification,” in *Proceedings of the 15th International Conference on Machine Learning*, Madison, WI, Jul. 1998, pp. 350–8

14. T. M. Mitchell, Machine Learning. New York: McGraw-Hill, 1997.

15. A. H. van't Hoog *et al.*, A systematic review of the sensitivity and specificity of symptom and chest-radiography screening for active pulmonary tuberculosis in HIV-negative persons and persons with unknown HIV status World Health Org., 2013.

16. Jaime Melendez*, Bram van Ginneken, Pragnya Maduskar, Rick H. H. M. Philipsen, Helen Ayles, and Clara I. Sánchez "On Combining Multiple-Instance Learning and Active Learning for Computer-Aided Detection of Tuberculosis" IEEE transactions on medical imaging, vol. 35, no. 4, April 2016.

17. J. Melendez et al., “A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest X-rays,” IEEE Trans. Med. Imag., vol. 34, no. 1, pp. 179–192, Jan. 2015.

18. S. Andrews, I. Tsochantaridis, and T. Hofmann, “Support vector machines for multiple instance learning,” Adv. Neural Inf. Process. Syst., vol. 15, pp. 561–568, 2003.

19. Saket Anand, Sushil Mittal, Oncel Tuzel and Peter Meer," Semi-Supervised Kernel Mean Shift Clustering" IEEE

Transactions on Pattern Analysis and Machine Intelligence, **Page(s):** 1201 - 1215

20. C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, pp. 273–297, 1995.

21. B. Settles, Active learning literature survey Univ. Wiscosin- madison, Comput. Sci tech Rep. 1648, 2009.

22. S. Tong and D. Koller, “Support vector machine active learning with applications to text classification,” *J. Mach. Learn. Res.*, vol. 2, pp.45–66, 2001.

23. M.-F. Balcan, A. Broder, and T. Zhang, “Margin based active learning,” in *Proc. COLT*, 2007, pp. 35–50.

24. X. Zhu, P. Zhang, X. Lin, and Y. Shi, “Active learning from data streams,” in *Proceedings of the 7th IEEE International Conference on Data Mining*, Omaha, NE, Oct. 2007, pp. 757–62

25. S. Dasgupta and D. Hsu, “Hierarchical sampling for active learning,” in *Proc. ICML*, 2008, pp. 208–215.

26. E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, “Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach,” *Biometrics*, vol. 44, pp. 837–845, 1988.

27. R. Dawson *et al.*, “Chest radiograph reading and recording system: Evaluation for tuberculosis screening in patients with advanced HIV,” *Int. J. Tuberc. Lung Dis.*, vol. 14, pp. 52–58, 2010.

28. D. Comaniciu and P. Meer, “Mean shift: A robust approach toward feature space analysis,” *IEEE Trans. Pattern Anal.*, vol. 24, no. 5, pp. 603–619, May 2002.

29. S.-J. Huang, R. Jin, and Z.-H. Zhou, “Active learning by querying informative and representative examples,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 10, pp. 1936–1949, Oct. 2014.

30. K. Brinker, “Incorporating diversity in active learning with support vector machines,” in *Proceedings of the 12th* International Conference on Machine Learning (ICML'03), Washington, DC, Aug. 2003, pp. 59-66.

31. A. Mucherino, P. J. Papajorgji, and P. M. Pardalos, “k-nearest neighbor classification,” in *Data Mining in Agriculture*, Vol. 34, Springer, New York, NY: Sep. 2009, pp. 83–106.

32. B. Settles, M. Craven, and S. Ray, “Multiple-instance active learning,” *Adv. Neural Inf. Process. Syst.*, vol. 20, pp. 1289–1296, 2008.

33. D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. Machine Learning, 15(2):201–221, 1994.