Active Learning for Microarray based Leukemia Classification

Kecheng Zhu
College of Engineering, Virginia Tech, Blacksburg, VA
kechengz@vt.edu

ABSTRACT

In machine learning, data labeling is assumed to be easy and cheap. However, in real word cases especially clinical field, data sets are rare and expensive to obtain. Active learning is an approach that can query the most informative data for the training. This leads to an alternative to deal with the concern mentioned above. The Sampling method is one of the key parts in active learning because it minimizes the training cost of the classifier. By different query method, models with considerable difference could be produced. The difference in model could lead to significant difference in training cost and final accuracy outcome. The approaches that were used to in this experiment is uncertainty sampling, diversity sampling and query by committee. In the experiment, active learning is applied on the microarray data with improving results. The classification on two types leukemia (acute myeloid leukemia and acute lymophoblastic leukemia) indicates a boost in accuracy with the same number of samples compared to passive machine learning. The experiments leads to the conclusion that with small number of samples with randomness in the field of leukemia classification, active learning produce an more active model. Additionally, active learning with query by committee finds the most informative sample with fewest trials.

CCS CONCEPTS

• **Computing methodologies**; • **Machine learning**; • **Learning settings**; • **Active learning settings**;

KEYWORDS

Active learning, Query Method, Pool-based, Uncertainty sampling, Query by committee

ACM Reference Format:
Kecheng Zhu. 2021. Active Learning for Microarray based Leukemia Classification. In *2021 8th International Conference on Biomedical and Bioinformatics Engineering (ICBBE '21), November 12–15, 2021, Kyoto, Japan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3502871.3502884

1 INTRODUCTION

Human genes generally keep the inherent information. A genetic disease is a disease caused by mutation. Genetic diseases are caused by changes in genetic disposition or controlled by pathogenic genes. Leukemia is a type of bone marrow cancer due to a genetic mutation. The recent society changes increase the probability of leukemia and most cases arise without any detectable symptoms.[5] At the early stage of leukemia, normal cells turn into cancerous cells which do not perform their job and gradually replace other cells.[1] The traditional way of determining the Deoxyribonucleic acid(DNA) mutation is complicated and requires a well-educated specialist.[2] With the new DNA micro-array analysis technology, it provides a way to determine whether an individual has a genetic mutation by analyzing large numbers of genes in a human body.[3] Therefore have a way to predict the risk of having leukemia from the result of the micro-array analysis.

Tradition leukemia prediction is a difficult job for manual work due to the large data set produced by the micro-array analysis, but with the development of machine learning algorithms and the increasing number of data that could be generated digitally, more and more work could be completed easier with computer support-machine learning algorithm. The previous approach to this problem is passive machine algorithms. Although passive machine learning have achieved a lot success in bio-informatics researches in other fields, the required clinical labeled data are limited, expensive and time-consuming to collect. This is because cases of leukemia are relatively rare and the raw patient data need well-educated specialists to put a large amount of effort to label. Therefore, with the limited samples, active learning is supposed to produce a model with higher accuracy. [4] Active learning, unlike passive machine learning that randomly selects the data to train the model. Active learning finds cases that it confused the most, and then through human interaction to confirm or get cases from the data set. Then semi supervised learning or supervised learning is carried out according to the obtained data to gradually improve the accuracy of the model. Instead of requiring all possible data to produce an accurate model, active learning only needs some samples that the model have the most confusion. With the help of Active Learning, the cost of labeling and collecting data can be significantly reduced.

2 METHODOLOGY

2.1 Active learning

In current scenarios of supervised learning, a large set of completely labeled data could not be available. Because the labeling process is too expensive or currently available data for training a model is considered to be less than necessary for an accurate model. Therefore, an alternative to that would be Active learning. Active learning will train a model by selecting the most useful sample from unlabeled data set and submit it to Oracle for annotation. It aims to reduce the amount of labeling cost while keeping the outcome same. Active learning can be classified into the following based on applied situation: membership query synthesis, stream-based selective sampling and pool based active learning [12]. Membership query synthesis is a way allowing the model to ask for samples in all unlabeled samples, including sample generated by the model. Stream-based selective is a way allowing the model to have the

ICBBE '21, November 12–15, 2021, Kyoto, Japan

ACM ISBN 978-1-4503-8507-7/21/11.
https://doi.org/10.1145/3502871.3502884

	Gene Description	Gene Accession Number	1	call	2	call.1	3	call.2	4	call.3	...	29	call.33	30	call.34	31	call.35	32	call.36	33	call.37
0	AFFX-BioB-5_at (endogenous control)	AFFX-BioB-5_at	-214	A	-139	A	-76	A	-135	A	...	15	A	-318	A	-32	A	-124	A	-135	A
1	AFFX-BioB-M_at (endogenous control)	AFFX-BioB-M_at	-153	A	-73	A	-49	A	-114	A	...	-114	A	-192	A	-49	A	-79	A	-186	A
2	AFFX-BioB-3_at (endogenous control)	AFFX-BioB-3_at	-58	A	-1	A	-307	A	265	A	...	2	A	-95	A	49	A	-37	A	-70	A
3	AFFX-BioC-5_at (endogenous control)	AFFX-BioC-5_at	88	A	283	A	309	A	12	A	...	193	A	312	A	230	P	330	A	337	A
4	AFFX-BioC-3_at (endogenous control)	AFFX-BioC-3_at	-295	A	-264	A	-376	A	-419	A	...	-51	A	-139	A	-367	A	-188	A	-407	A

Figure 1: Raw data set

	0	1	2	3	4	5	6	7	8	9	...	7121	7122	7123	7124	7125	7126	7127	7128	patient	cancer
0	-214	-153	-58	88	-295	-558	199	-176	252	206	...	-125	389	-37	793	329	36	191	-37	1	ALL
1	-139	-73	-1	283	-264	-400	-330	-168	101	74	...	-36	442	-17	782	295	11	76	-14	2	ALL
2	-76	-49	-307	309	-376	-650	33	-367	206	-215	...	33	168	52	1138	777	41	228	-41	3	ALL
3	-135	-114	265	12	-419	-585	158	-253	49	31	...	218	174	-110	627	170	-50	126	-91	4	ALL
4	-106	-125	-76	168	-230	-284	4	-122	70	252	...	57	504	-26	250	314	14	56	-25	5	ALL

Figure 2: Data after prepossessing

ability to identify samples needed for labeling as it accepts an input stream. Pool-based active learning allows the model to evaluate the sample space and have the ability to query the most informative sample(s) for labeling (Figure 1, Figure 2).

2.2 Pool-based Active learning

Pool based Active learning is to start the training with a relatively small amount of data. A query method adds new data to the training set or ask for labels to the unlabeled data based on specified principles, e.g Uncertainty sampling, query by committee, diversity sampling. These principles are defined based on sample space evaluation and current model prediction evaluations. One of the evaluations used for find the most informative point is uncertainty sampling. Uncertainty sampling is to query points for sample that the current model has the lowest confidence [6]. It assumes that labeling the most confused sample would improve the model. It is a straight forward approach based on confusion matrix. But since the results are depend on a single model's approach. The sample point found could be miss leading and cause drop points in learning curve. Diversity sampling another approach which specifies a distance function. The distance function will find the most distinct sample to the current given samples. However because the most distinct point could be extreme values or samples with high variances, this approach could be miss-leading to the training process. Then it leads to an approach that is based on disagreements between confusion matrix of models - Query by committee. Query by committee build multiple models and choose points where they disagree the most. This approach can avoid the sensitivity of model choices. But it is a more expensive approach because multiple models need to be trained for evaluation. These query method is aim to find the most informative sample for training. Until a specific goal is reached the algorithm stops [8, 9].

The pool based active learning assumes that we have a pool of samples that is unlabeled. Label a few at the start to train the model. An infallible oracle, it labels the acquired unlabeled data that is always correct, gave labels to those unlabeled acquired samples. It aims to improve the accuracy of the model. The pseudo code indicates the pool-based active learning is shown below:

- DataPool = readData()
- datagiven = []; done = false
- While ~done:
- [p,u] = model(datagiven)
- Rounds.append(getaccuracy(p))
- If max (u)<giventhresholdvalue
- s = query(u)
- Datagiven.append(getData(s))
- Else
- Done = true

2.3 Feature selection

The microarray technologies uses DNA from all chromosomes and checks if targeted chromosome is presented. Therefore, a thousands of information would be generated each test. [7] In case to classify the result of microarray data, instead of using all gene expressions as feature. It needs to select a sample of gene expression from the whole sample space to analysis. Because the large number of feature would result in a poor generalization ability in the result model. In the experiment, a variance based principal components analysis

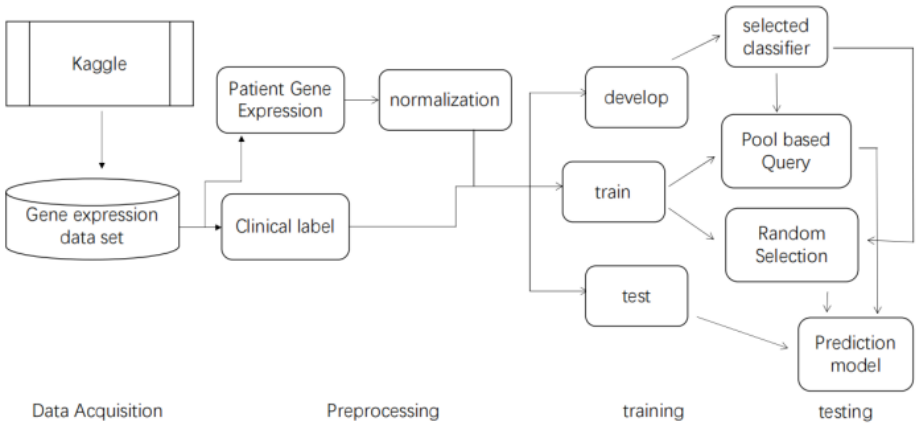

Figure 3: experiment architecture

(PCA) is used to reduce the number of sample gene expression to a number that is far lower than the original sample space. The number of features remaining is based on the following ration where the numerator is the averaged squared projection error PCA tries to minimize and the denominator is the total variation in data. In the experiment, this ratio is set to lower then 0.05 and obtain a reasonable number of features in the samples, 36 features with 95 percent of total explained variance (The contribution rate of each feature, sum is 1) were selected.

$$\frac{\sum \| x_{\mathrm{i}} - x_{\mathrm{iapprox}} \|^2}{\sum \| x_i \|^2} \tag{1}$$

3 EXPERIMENT ARCHITECTURE

3.1 Data set

The gene expression data set(Golub et al.) data set used was obtained from "kaggle" https://www.kaggle.com/crawford/gene-expression. The data set is about gene expression from Bone Marrow and Peripheral Blood and used for classifying different cases of cancer via DNA microarray with 7129 gene expression and 72 samples. The samples included two types of leukemia: 47 cases of acute myeloid leukemia (AML) and 25 cases of acute lymophoblastic leukemia(ALL). The Intensity values have been re-scaled such that overall intensities for each chip result are equivalent [11]. Because all data and results are existed. The experiment is going to be done in an retrospectively way. The results of data would be hidden to the model and only reveal the results when asked. In this way to approximate the situation of a pool-based active learning.

3.2 Experiment set up

The figure 3 presents the experiment architecture for classification of leukemia. It includes the following phases: data acquisition, prepossessing, training and testing.

The prepossessing part is to combine unlabeled data from different files and label it for classification task. First, unlabeled clinical microarray results from different files were combined and duplicates were removed. Then unnecessary descriptions to gene expression were removed, e.g support calls for intensity values of presence of gene, gene description. After that with the provided clinical result, each patient were labeled their type of leukemia.

In training part, the data set were normalized to remove extreme values to improve the result of the classification process outcome. Then the data was split into development, train and testing set. Initially, several classification model were selected to test on the development set to find a suitable model based on receiver operating curve (ROC). ROC is a way to examine the sensitivity and specificity of a model on a specific task. The horizontal axis indicates the number of false positive rate(specificity) and the vertical axis indicates the True positive rate(sensitivity). Therefore, the closer the ROC curve is to the upper left corner, the better the performance of the classifier, which means that the classifier has a very high true positive rate while the false positive rate is very low. By examine the ROC curve, linear SVM classifiers were selected. After that, two ways of sample query is employed: random sampling (passive machine learning), pool-based query(active machine learning). The pool-based query is to provide the training set with no or few labels. And allow the model to have limited access to an oracle (a single infallible oracle) that can query to get a label for a specific object to train the model [13].

In test phases, the learning curve produce from random sampling and pool-based query were compared for linear SVM classifier. And both model were taken to predict the test samples, the accuracy, recall, and F1 score are taken into account to evaluate the final outcomes.

3.3 Experiment Results

As mentioned in the methodology, the figures indicates different Active learning sampling method compare to Passive machine learning (random sampling). The learning curve reveals a significant increase accuracy in the final outcome model. In random sampling, the model gets an average of 62% accuracy and 55% recall. In Active machine learning gets an average accuracy rate of 83% and a 82% recall.

As the results indicates, the random sampling's learning curve fluctuates and even have drops in accuracy when the budget is used up compared. There is some points where the accuracy for the

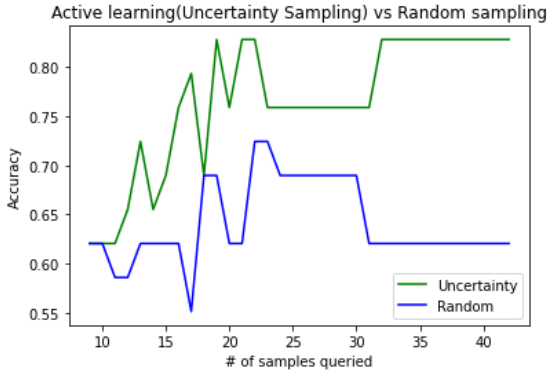

Figure 4: uncertainty sampling vs random sampling learning curve

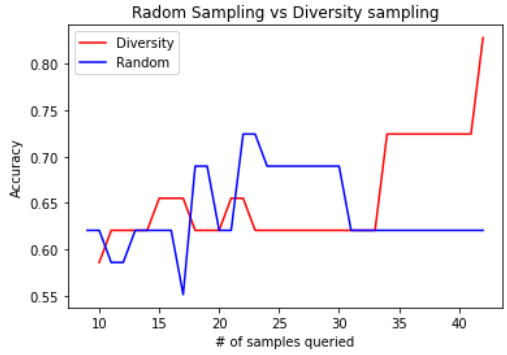

Figure 5: Diversity sampling vs random sampling learning curve

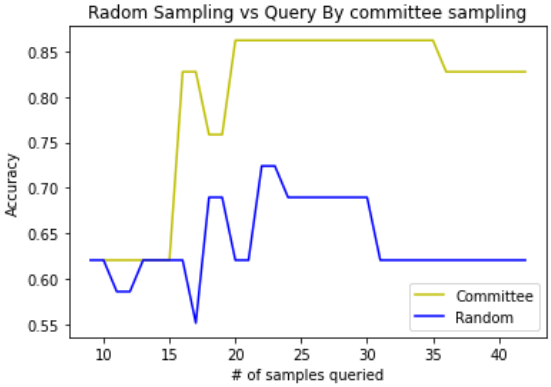

Figure 6: query by committee vs random sampling

random sampling's prediction is relatively high because of the size of the testing and training sample space is small which could result in drops in the final model. In figure 4, the uncertainty learning curve indicates a boost at first few queries.

As method in methodology, the confusion matrix helped the training process by finding the most informative sample. However, there are significant fluctuation during the boost process which indicates the take samples leads to a significant change in the model. In figure 5, the diversity sampling is based on a distance function that find the farthest point to current samples. The samples did not learn quite much until last few samples were queried. This indicates that the diversity sampling method did not help the training process for the most informative points but it is finding extreme points. In figure 6, the query by committee sampling's learning curve indicates a significant boost and is followed by a relative horizontal curve. This means that during the boost era, all the informative points were precisely selected by the query method.

4 CONCLUSION

The contribution of this research is compared the active machine learning to passive machine learning in the situation of microarray based leukemia classification. Moreover, compared different active machine learning sampling method in machine learning for micro array analysis.

The introduce of active machine learning allow the training process to acquire more efficient training data to improve the model. This means that producing the same accuracy for a model with less labeling effort. The methodology is built on a data set that contains gene expressions for all kinds of gene analysis and can be used for other kinds of genetic disease filtering [14, 15].

In this experiment, there are several issues could be improved to achieve a better result. First of all, the sampling method query by committee was based on a naive approach, it is based on disagreements between models. However, situations of existing extreme prediction values between models could make the query less informative or even miss-leading to the model. This problem could be improved by adding a voting algorithm, e.g voting entropy, average Kullback-Leibler Divergence [16]. Second, in the uncertainty sampling method, only criteria used is least confidence. Marginal sampling and entropy approach did not take into consideration of this experiment. To this experiment, classifying AML and ALL is a binary classification problem. This means that the Marginal sampling should be equivalent to least confidence. But Entropy take consider of all possible predictions to a given value, which could result in a different result [10].

To resolved issue mentioned earlier, some follow up studies were planned. First of all, an improved uncertainty sampling method with diversity maximization [17] would be implemented and compared to the query by committee sampling method. Second, more complicated sampling method, e.g expected model change, other forms of query by committee, would compared to the current sampling strategy. Last But not the least, after comparing these sampling methods, the outcome methods would used to train for other classification tasks based on identifying generic diseases from microarray analysis.

To sum up, In active learning, the key is to find the most informative sample candidates and ask for labeling. The way to find the most informative sample is the query method. The query method could base on the outcome of a single model or several models. By selecting the most informative samples could reduce labeling costs and improve learning efficiency. Moreover, for those generative

models, cased generated by the model could help the training cases where clinical data is inefficient to get a model with high precision.

REFERENCES

[1] Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression, Science 286:531-537. (1999). Published: 1999.10.14, T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander.

[2] Mallick, P.K., Mohapatra, S.K., Chae, GS. *et al.* Convergent learning–based model for leukemia classification from gene expression. Pers Ubiquit Comput (2020). https://doi.org/10.1007/s00779-020-01467-3

[3] Ratley A, Minj J, Patre P (2020) Leukemia disease detection and classification using machine learning approaches: a review. In: 2020 First International Conference on Power, Control and Computing Technologies (ICPC2T). IEEE, pp 161–165.

[4] D. Vogiatzis, N. Tsapatsoulis,Active learning for microarray data,International Journal of Approximate Reasoning,Volume 47, Issue 1,2008,Pages 85-96,ISSN 0888-613X, https://doi.org/10.1016/j.ijar.2007.03.009..

[5] Abelson, S., Collord, G., Ng, S.W.K. et al. Prediction of acute myeloid leukaemia risk in healthy individuals. Nature 559, 400–404 (2018). https://doi.org/10.1038/s41586-018-0317-6.

[6] McClintick, JeanetteN, and HowardJ Edenberg. BMC Bioinformatics, vol. 7, no. 1, 2006, p. 49., doi:10.1186/1471-2105-7-49. .

[7] Ronan, Anne.Chromosome microarray analysis: A soothing guide[J].Journal of Paediatrics and Child Health,2018.

[8] Liu, Dongjiang, Liu, Yanbi.An active learning algorithm for multi-class classification[J].Pattern Analysis and Applications,2018

[9] Journal of the American Medical Informatics Association, Volume 21, Issue 5, September 2014, Pages 893–901, https://doi.org/10.1136/amiajnl-2013-002516

[10] Krishnapuram, Balaji, *et al.* Cost-Sensitive Machine Learning. CRC Press, 2012.

[11] Aggarwal, Charu C., *et al.* "Active learning: A survey." Data Classification: Algorithms and Applications. CRC Press, 2014. 571-605.

[12] Figueroa RL Zeng-Treitler Q Ngo LH *et al.* . Active learning for clinical text classification: is it better than random sampling? J Am Med Inform Assoc 2012;19:809–16.

[13] Nguyen D Patrick J . Reverse active learning for optimising information extraction training production. In: Thielscher Michael Zhang Dongmo, eds. AI 2012: advances in artificial intelligence. Springer Berlin Heidelberg, 2012:445–56.

[14] Corne, David W., and Gary B. Fogel. "An Introduction to Bioinformatics for Computer Scientists." Evolutionary Computation in Bioinformatics, 2003, pp. 3–18., doi:10.1016/b978-155860797-2/50003-2.

[15] Singh, Dilbag, and Jasjit Singh Samagh. "A COMPREHENSIVE REVIEW OF HEART DISEASE PREDICTION USING MACHINE LEARNING." Journal of Critical Reviews, vol. 7, no. 12, 2020, doi:10.31838/jcr.07.12.54.

[16] Active learning, B Settles, 2012

[17] IJCV 2014 " Multi-Class Active Learning by Uncertainty Sampling with Diversity Maximization"