Active Bagging Ensemble Selection

Vangjel Kazllarof, Sotiris Kotsiantis

Active Bagging Ensemble Selection

Vangjel Kazllarof[1][0000-0003-4545-3866] and Sotiris Kotsiantis[2][0000-0002-2247-3082]

[1] University of Patras, 26504 Rio Achaia, Greece
[2] University of Patras, 26504 Rio Achaia, Greece

Abstract. As technology progresses with more and more data collected, the need of finding the appropriate label for them increases. However, many times the labeling process is a very difficult or/and expensive task and in most cases a help of an expert or expensive equipment is needed. For this reason the need of labeling only the most appropriate instances rises. Active Learning techniques can accomplish this by querying only those instances that a trained model finds the greatest amount of information and providing them to a human expert in order to label them. Combining these techniques with a fast ensemble classifier, a very performant in terms of classification accuracy schema can emerge where a trained model in a small amount of labeled instances can grow by adding only the most informative instances from a much greater pool of unlabeled instances. In this paper, we will propose such a schema using Bagging Ensemble Selection that uses REPTree as base classifier under Active Learning techniques and we will compare it to four well-known ensemble classifiers under the same techniques on 61 real world datasets.

Keywords: Active Learning, Ensemble Classifiers, Algorithms.

1 Introduction

Nowadays large amount of data are collected from various sources. However, not all of this data comes with a classified label because it might be quite difficult, time consuming [1] or/and expensive [2] to find it. Many times the labeling process is bounded by a budget plan allowing only a small amount of the unlabeled data for annotating. For these reasons creating accurate models with the use of as least as possible data becomes essential. This rises the need of finding a way to discard the data that provide no useful information and find the label for only the ones that will improve the initial model. A field in Machine Learning called Weakly Supervised Learning has emerged that tackles the problems described above, with Active Learning (AL) [3] playing a leading role in the field. The AL process is utilized by performing queries to the unlabeled data in order to find those instances that provide the most amount of information. Essential fact to create accurate models under AL schema is the use of robust and accurate ensemble classifiers with fast and accurate base classifiers.

In this work we will propose an AL schema with the use of Bagging Ensemble Selection (BES) that uses Reduced Error Pruning Tree (REPTree) [4] as base classifier and we will compare it with other well-known ensemble classifiers under the same AL

schema proving that the proposed one can outperform the rest in terms of classification accuracy, especially when the portion of initial labeled data is very small compared to the unlabeled one.

In the following section a brief introduction of the AL theory and the BES with the use of REPTree as base classifier will be presented along with the related work. In section 3, we will demonstrate the proposed algorithm. The experimental procedure and the results of our experiments will be described in section 4 following with the conclusions and the future work in section 5.

2 Related Work

To begin with, a number of techniques in querying unlabeled instances have been developed and they are called Scenarios. Membership Query Synthesis [5] is one of them in which the model generates instances de novo and then queries them. Another is the Stream Based Selective Sampling [6] in which unlabeled instances come one-by-one from an input stream and the model decides whether will be queried or discarded. The last one is Pool Based [7] scenario, probably the most popular one, in which assumes that there is a small pool of labeled instances (L) and a much larger pool of unlabeled ones (U). The model rates all the instances of U and selects those with the greatest value to query.

For measuring the informativeness of the unlabeled instances, several Query Strategies (QS) have been developed [3]. Uncertainty Sampling [7] is one of the most known strategies in which the model rates the instances based on the uncertainty level. The selected instances for querying are the one that the model is most uncertain.

AL methods with the use of ensemble classifiers have been widely used in both experimental and real life problems. In Gesture Recognition area, algorithms that use AL with the use of a number of ensemble methods and Random Forest have been proposed in [8] and [9] respectively. In both works a good classification accuracy has been achieved by using only a small subset of the whole dataset. Continuing in the Image Recognition field, an ensemble of Convolutional Neural Networks has been proposed in [10] under an AL schema for the uncertainty estimation of unlabeled data. It was compared against Monte Carlo Dropout and geometric approaches showing that the proposed algorithm outperformed its rivals in classification image datasets used in the literature and in real-world images for diabetic retinopathy detection. A study on the Data Stream Mining have been demonstrated in [11] where Active Online Ensembles is proposed in which accurate models are built with the use small-sized ensembles. In the study online Bagging and online Boosting that use Hoeffding tree and k-Nearest Neighbors as base classifiers have been extended with an AL component and showed great results in classification accuracy with much smaller ensemble size against traditional online Bagging and online Boosting contributing in the training performance as well.

Moreover, algorithms that combine AL and ensemble trees have been proposed in [12] and [13] in which Rotation Forest and Logitboost with the use of M5P as base

classifier have been exploited respectively, showing good results in classification accuracy against other simple and ensemble classifiers under the same AL schemas on real-world datasets. Similar approach is implemented in the current work, adding BES with the use of REPTree base classifier in the pool of accurate ensembles under AL schema.

3 Active Learning BESTrees

3.1 Bagging Ensemble Selection

For the ensemble strategy, an extension of Ensemble Selection (ES) [14] that uses a library of base learners to construct the ensemble one is chosen in this work. It uses a construction strategy in order to extract a well performing subset of models that are trained with the base learners. Forward step-wise selection strategy was proposed in [14] that started with an empty ensemble and it was gradually populated by the models that maximize the ensemble's performance on a hillclimb set, until all models are examined.

Although ES has shown great advantages compared to other ensemble strategies, many times it has suffered from overfitting on the hillclimb set, hitting the performance of the model on the test set. To overcome this, BES [15] was proposed that in its simplest implementation, ES is treated as a base classifier for bagging.

Moreover, an extension of BES, named BES-out-of-bag (BES-OOB), was proposed in [15] that uses all of the bootstrap sample in order to generate the models and the respective out-of-bag (OOB) instances as the hillclimb set. Then, base classifiers were trained in the bootstrap sample and ensemble selection was applied to them according to their performance in the OOB sample.

The implementation of BES-OOB method with the use of REPTree (BESTrees) as base classifier was compared with other ensemble strategies in [16] showing better results against Stochastic Gradient Boosting and Bagging and being comparable against an ensemble that combines both Bagging and Stochastic Gradient Boosting.

Although BES has shown great performance, it has not been widely used in the literature. BESTrees was found to be used for Sentiment Classification in [17] resulting, however, worse results than Random Forest, on a dataset produced by tweets. More promising results was shown on a lightweight extension of BESTrees in [18] where BES classification was used for Activity Recognition on a smart home system using mobile phones and iBeacons. It was compared against other existing lightweight algorithms outperforming them in both classification accuracy and efficiency in terms of hardware resources.

3.2 Proposed Algorithm

First of all, the Labeled Ratio (R) is defined as the percentage of the size of L compared to the sum of the sizes of L and U and it is described by the following formula:

$$R = \frac{Size(L)}{Size(L) + Size(U)} * 100 \quad (1)$$

Starting now with L sized with a small R, the AL procedure creates an initial model and then uses it to rate the instances of U, given the selected QS. After finding the most informative instances, the next step is to provide them to a Human Expert (HE) one-by-one or in batches of fixed or dynamic size, depending of the size of the initial L. The HE is an ideal labeler in this work that always annotates correctly the provided instances. Then, the new labeled instances are removed from U and added to L, ending an AL circle [3].

In the next step, the new formatted L is trained again and this procedure goes on until the stopping criterion is met. For the stopping criterion we used the max iterations method by stopping the whole process after 15 iterations. The output of the algorithm is the trained model on L formatted in the last iteration of the algorithm. This model then is used to predict the test set.

The chosen ensemble classifier is the BESTrees as described above and under AL schema the Active Learning BESTrees (ALBESTrees) emerges and it is described in the following algorithm:

```
Algorithm ALBESTrees:
  HE: Human Expert (Oracle);
  CL: Classifier (BESTrees);
  MI: Max Iterations (15);
  QS: Query Strategy (Uncertainty Sampling);
  I:  Current Iteration (0);
  L:  Initial Labeled Set;
  U:  Initial Unlabeled Set;
  R:  Labeled Ratio;
  T:  Test Set;
  B:  Batch Size (Ceil((Size(L)+Size(U)) * R / MI));
  TR: Top Rated Instances;
  Begin:
    Train CL in L;
    While I < MI:
      Assign uncertainty values to U using QS(CL);
      Add B top rated instances in TR;
      Remove TR from U;
      Ask HE for labeling instances in TR;
      Merge TR in L;
      Empty TR;
      Retrain CL in new L;
      Assign I+1 in I;
    End
    Predict labels of T using last trained CL;
  End
End
```

For each iteration we used the batch mode sampling (B) where the top rated instances are selected in batches from the U set. The size of the batches is different for each dataset and it is calculated with the following equation:

$$B = \left\lceil \frac{\left(Size(L) + Size(U)\right) * R}{Max\ Iterations} \right\rceil \quad (2)$$

This way, we eliminate the size dependency of the datasets compared to the iterations. This way, the size of L in the last iteration will be always doubled from starting the size of the procedure.

4 Experiments and results

4.1 Datasets

For our experiments we selected 61 real-world datasets from UCI repository [19]. The datasets consist of 27 binary and 34 multiclass classification problems. They have a great variety of sizes with the smallest one to have only 208 (sonar) instances while the largest 67557 instances (connect-4). The number of classes vary from 2 classes for binary datasets to a maximum of 28 classes for multiclass (abalone). As for the number attributes, all datasets have a range of 2 (banana) to 90 (movement_libras) different attributes where many of them are constructed form only numerical or categorical types while other from both.

For the imbalance factor, the binary datasets are almost all balanced with only one exception (coil200) being quite imbalance with almost 6% of instances classified in the minority class. On the other hand, many multiclass datasets have a great imbalance factor with the instances classified to the minority class less than 5%.

4.2 Experimental procedure

The AL schema starts with Pool Based scenario with the use of a small L and big U as starting sets. For R we selected four different values to experiment with and these are 5%, 10%, 15% and 20% in order to examine the behavior of the algorithm in both small and big initial L. The model in the last iteration is trained to an L doubled in size being 10%, 20%, 30% and 40% of the whole dataset for each R value respectively.

In order to rate the instances in the U set, we used Uncertainty Sampling (UncSamp) with three different methods in order to measure the uncertainty. The first is Entropy in which the top rated instances are the one that increase the entropy of the model, the second is Least Confidence that rates the instances in favor of the one that the model has the least confidence and the last one is Smallest Margin that aims to select the instances that have the smallest margin between the two most probable classes that the model decides. To compare with, we also run the experiments with Random Sampling that selects the instances from U in a random manner.

For rivals we selected 3 well-known ensemble techniques with the use of 3 different tree classifiers as base classifier. The first one is Bagging technique with base classifiers the REPTree, that is used in the proposed algorithm as well, and J48 that is a Java

implementation of the C4.5 decision tree. The next two ensemble techniques belong to the Boosting family and these are LogitBoost [20] and AdaBoostM1 [21] with the use of Decision Stump as base classifier.

In the training process we used 3-fold validation that separates the train set in three folds and uses 2/3 for training and the rest 1/3 for testing, repeating the process three times until all folds are tested. The classification accuracy of the model is the average of results of each testing phase.

It is worth mentioning that every experiment ran 3 times for every dataset, QS, R and classifier and the final result is the average of the results of each experiment. For software we used WEKA [4] with the use of JCLAL [22] framework for the AL setup and for each classifier we used the default parameters provided by the software.

4.3 Results

In Table 1 and Table 2 we demonstrate the results of the classification accuracy of each algorithm for both binary and multiclass datasets respectively annotating with bold text the best classification accuracy for each dataset. Due to lack of space, we only included the results of classification accuracy for R=5% using the Uncertainty Sampling with Entropy method against the rest of the selected ensemble methods. The rest of the results can be found in the link:
http://ml.math.upatras.gr/wp-content/uploads/2021/03/Active-BES-Results.zip.

Moreover, we calculated the total winnings of each algorithm in terms of classification accuracy for every initial R and every query strategy method for both binary and multiclass datasets. The proposed algorithm outperformed its rivals in all of the uncertainty sampling metrics and initial labeled ratios with a total of 574 winnings followed by Bagging(J48) with 127 winnings. It is worth mentioning that most winnings of the proposed algorithm are accomplished for R = 5% proving that it is a better choice when there are very few labeled instances compared to the unlabeled one.

For the statistical analysis of the performance of the algorithms, we used non-parametrical Friedman tests that examines if the null hypothesis of the similarity of the algorithms that it compares holds [23]. In Fig. 1, we demonstrate these results comparing all algorithms with the use of Uncertainty Sampling and Random Sampling query strategies using violin plot.

From the results it is shown that the proposed algorithm has the best ranking in all cases compared to its rivals. Moreover, Uncertainty Sampling shows better rankings compared to Random Sampling proving that AL techniques can be very beneficial in order to choose the most appropriate instances.

5 Conclusions and Future Work

In this work we proposed BES with the use of REPTree as base classifier under AL schema. We compared it with other well-known ensemble methods on both binary and multiclass classification datasets under the same AL schema. From the results it is shown that the proposed algorithm outperforms its rivals with statistical significance

according to non-parametrical Friedman tests, making it a great candidate for problems where acquiring labels of unlabeled data is constrained by a budget plan.

For future work, we will compare the proposed algorithm with other state-of-the-art ensemble algorithms like Random Forest, Rotation Forest and Neural Networks Ensembles in both classification accuracy and training efficiency. Moreover, we will experiment with Noisy Oracles where HE has an error ratio on its annotations, in order to simulate more real life scenarios [24].

Table 1. Classification accuracy of the ALBESTrees against the selected ensemble methods for binary datasets under UncSamp(Entropy) strategy for R = 5%

Dataset	BES (REPTree)	Bagging (REPTree)	Bagging (J48)	Logitboost (Decision St.)	Adaboost (Decision St.)
banana	70.81	82.10	83.35	**84.48**	59.25
bands	**56.90**	44.74	47.21	49.32	55.14
chess	91.94	93.19	**95.78**	90.00	84.38
coil2000	**94.03**	93.96	93.49	92.30	**94.03**
credit-a	**85.94**	71.40	81.06	72.75	84.88
credit-g	**71.57**	70.13	69.53	68.53	70.63
german	**71.57**	70.10	68.73	68.37	70.47
heart-statlog	**72.59**	55.56	60.25	69.14	70.49
housevotes	**96.98**	95.25	94.67	91.82	94.82
ionosphere	**87.84**	56.89	62.20	69.92	74.45
kr-vs-kp	92.09	94.50	**95.58**	90.39	86.57
magic	**85.59**	83.26	82.88	81.73	77.14
mammo-graphic	**83.41**	70.68	79.52	74.66	79.92
monk-2	**98.07**	97.22	97.22	90.48	95.76
mushroom	**99.94**	99.74	99.36	99.41	97.58
phoneme	**83.78**	79.16	80.47	78.26	72.25
pima	**73.39**	67.97	69.01	69.84	70.66
ring	**95.58**	85.67	87.62	82.30	49.51
sonar	**62.85**	45.53	52.57	59.48	60.73
spambase	**92.94**	88.80	89.66	85.08	83.76
spectfheart	**76.78**	43.07	49.44	59.70	69.91
tic-tac-toe	71.40	67.88	67.40	**75.01**	69.07
titanic	71.20	77.65	**78.27**	77.15	77.66
twonorm	**95.82**	84.77	86.36	85.82	84.81
vote	90.50	78.08	88.35	91.56	**92.72**
wdbc	**94.43**	85.18	85.76	89.87	94.38
wisconsin	**96.88**	89.95	92.97	92.02	92.97

Table 2. Classification accuracy of the ALBESTrees against the selected ensemble methods for multiclass datasets under UncSamp(Entropy) strategy for R = 5%

Dataset	BES (REPTree)	Bagging (REPTree)	Bagging (J48)	Logitboost (Decision St.)	Adaboost (Decision St.)
abalone	**24.91**	21.88	19.94	18.73	16.73
anneal	**90.05**	84.11	85.15	89.03	77.02
audiology	29.80	33.04	**49.10**	38.91	33.90
balance-scale	73.92	62.61	71.84	**74.96**	49.81
balance	74.93	69.17	66.98	**76.28**	55.90
car	**83.82**	74.19	78.34	80.56	70.79
cleveland	**56.57**	53.87	53.87	51.70	53.98
connect-4	69.41	71.71	**72.67**	70.24	65.83
dermatology	**93.48**	75.69	86.31	76.17	48.70
ecoli	**74.80**	68.35	68.15	63.99	62.00
flare	69.23	**70.04**	68.73	68.49	53.47
kr-vs-kp	**36.34**	27.27	33.56	35.84	10.04
led7digit	**54.73**	41.59	47.00	48.78	14.94
letter	**80.84**	62.47	68.15	71.02	6.91
marketing	**31.53**	28.56	26.59	25.89	18.64
move-ment_libras	**35.74**	20.74	24.35	27.24	10.93
newthyroid	**81.41**	81.11	78.79	80.70	81.26
nursery	88.79	89.17	**90.74**	90.41	64.54
optdigits	**91.48**	79.94	81.60	78.35	18.74
page-blocks	**96.91**	95.74	96.06	95.04	92.60
penbased	**96.31**	89.92	91.95	89.72	20.52
primary-tumor	25.86	25.17	24.39	**28.81**	25.86
satimage	**86.23**	80.95	81.66	73.37	33.63
segment	**93.67**	88.66	89.25	85.34	28.51
shuttle	**99.96**	99.90	**99.96**	99.83	84.23
soybean	**64.13**	20.45	43.88	60.24	13.47
texture	**92.28**	81.48	84.93	83.45	16.08
thyroid	99.11	99.30	**99.36**	96.62	96.87
vehicle	**62.37**	48.35	52.44	56.34	26.04
vowel	**43.30**	10.91	39.66	35.93	14.14
waveform-5000	**82.03**	75.11	75.00	70.28	55.37
winequality-red	**49.86**	45.03	46.32	40.70	42.21
winequality-white	**50.52**	46.86	47.39	40.94	31.19
yeast	**49.10**	45.15	47.15	41.40	21.29

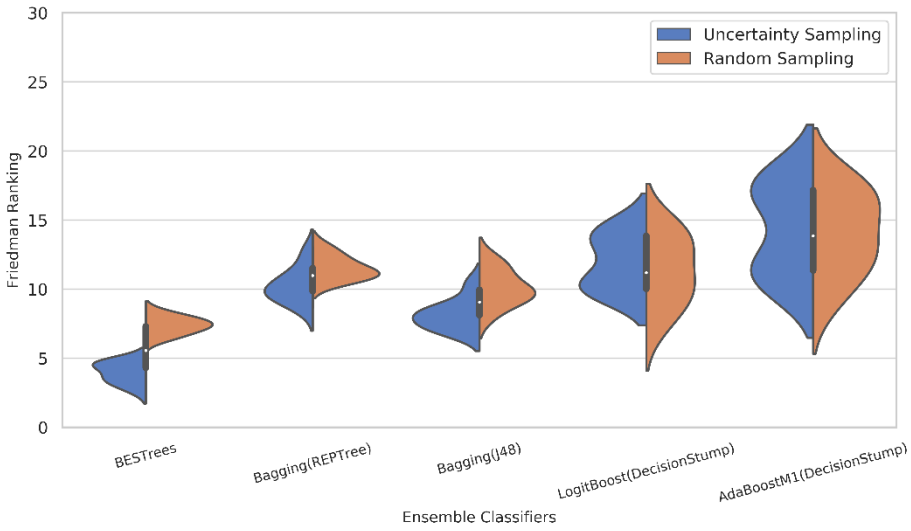

Fig. 1. Violin plot presenting the distribution of Friedman ranking of the ensemble classifiers comparing the selected query strategy against random sampling

References

1. Zhu, X., Lafferty, J., Ghahramani, Z.: Combining Active Learning and Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. In: ICML 2003 Workshow on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining. pp. 58–65 (2003)
2. Settles, B., Craven, M., Friedl, L.: Active Learning with Real Annotation Costs. In: Proceedings of the NIPS Workshop on Cost-Sensitive Learning. pp. 1–10 (2008)
3. Settles, B.: Active learning. Synth. Lect. Artif. Intell. Mach. Learn. 18, 1–111 (2012). https://doi.org/10.2200/S00429ED1V01Y201207AIM018
4. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. SIGKDD Explor. 11, (2009)
5. Angluin, D.: Queries and Concept Learning. Mach. Learn. 2, 319–342 (1987)
6. Cohn, D., Ladner, R., Waibel, A.: Improving generalization with active learning. Mach. Learn. 15, 201–221 (1994)
7. Lewis, D.D., Gale, W.A.: A Sequential Algorithm for Training Text Classifiers. In: Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 3–12. ACM/Springer (1994)
8. Schumacher, J., Sakič, D., Grumpe, A., Fink, G.A., Wöhler, C.: Active learning of ensemble classifiers for gesture recognition. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 498–507. Springer, Berlin, Heidelberg (2012)
9. Kazllarof, V., Kotsiantis, S., Karlos, S., Xenos, M.: Automated hand gesture recognition exploiting active learning methods. In: PCI 2017: Proceedings of the 21st Pan-Hellenic Conference on Informatics. pp. 1–6. Association for Computing Machinery (2017)

10. Beluch, W.H., Genewein, T., Nürnberger, A., Köhler, J.M.: The Power of Ensembles for Active Learning in Image Classification. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 9368–9377. IEEE Computer Society (2018)
11. Alabdulrahman, R., Viktor, H., Paquet, E.: An active learning approach for ensemble-based data stream mining. In: IC3K 2016 - Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. pp. 275–282. SciTePress (2016)
12. Kazllarof, V., Karlos, S., Kotsiantis, S.: Active learning Rotation Forest for multiclass classification. Comput. Intell. 35, 891–918 (2019). https://doi.org/10.1111/coin.12217
13. Kazllarof, V., Karlos, S., Kotsiantis, S.: Investigation of Combining Logitboost(M5P) under Active Learning Classification Tasks. Informatics. 7, 50 (2020). https://doi.org/10.3390/informatics7040050
14. Caruana, R., Niculescu-Mizil, A., Crew, G., Ksikes, A.: Ensemble Selection from Libraries of Models. In: Proceedings of the 21st International Conference on Machine Learning. pp. 137–144 (2004)
15. Sun, Q., Pfahringer, B.: Bagging Ensemble Selection. In: Proceedings of the 24th Australasian Joint Conference on Artificial Intelligence (AI'11). pp. 251–260. Springer, Perth (2011)
16. Sun, Q., Pfahringer, B.: Bagging Ensemble Selection for Regression. In: Proceedings of the 25th Australasian Joint Conference on Artificial Intelligence (AI'12). pp. 695–706. Springer, Sydney (2012)
17. Moreira, S., Filgueiras, J., Martins, B., Couto, F., Silva, M.J.: REACTION: A naive machine learning approach for sentiment classification. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 490–494. Association for Computational Linguistics, Atlanta (2013)
18. Alam, M.A.U., Pathak, N., Roy, N.: Mobeacon: An iBeacon-Assisted Smartphone-Based Real Time Activity Recognition Framework. In: Proceedings of the 12th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services. pp. 130–139. Institute for Computer Sciences, Social Informatics and Telecommunications Engineering (ICST) (2015)
19. Dua, D., Graff, C.: UCI Machine Learning Repository, http://archive.ics.uci.edu/ml
20. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: A statistical view of boosting. Ann. Stat. 28, 337–407 (2000). https://doi.org/10.1214/aos/1016218223
21. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. Proc. Thirteen. Int. Conf. Int. Conf. Mach. Learn. 148–156 (1996)
22. Reyes, O., Pérez, E., Fardoun, H.M., Ventura, S.: JCLAL: A Java Framework for Active Learning. J. Mach. Learn. Res. 17, 1–5 (2016). https://doi.org/10.5555/2946645.3007048
23. Eisinga, R., Heskes, T., Pelzer, B., Te Grotenhuis, M.: Exact p-values for pairwise comparison of Friedman rank sums, with application to comparing classifiers. BMC Bioinformatics. 18, 68 (2017). https://doi.org/10.1186/s12859-017-1486-2
24. Gupta, G., Sahu, A.K., Lin, W.Y.: Noisy Batch Active Learning with Deterministic Annealing. (2019)