Introduction to Scientific Working

Assignment 11

1 Paper-1

1.1 Reference

Blum, A. Mitchell, T.
Combining labeled and unlabeled data with co-training
Proceedings of the eleventh annual conference on Computational learning theory, 1998, 92-100

1.2 Claim

- In this article author has taken the task of classification of web pages
- Currently the web pages are classified by either supervised or unsupervised learning
- Author has suggested the method of combining the both methods in such a ratio of taking very few labelled data to a large set of unlabelled data with which unlabelled data are sampled and classified. The algorithm as follows
  - Consider L as event of labelled training example
  - U as a event of unlabeled training example
  - generate a pool of U' from the random sampling of U
  - Train the hypothesis h_1 using L for a small portion x_1 of x such that $x_1 \subset x$
  - Train the hypothesis h_2 using L for a small portion x_2 of x such that $x_2 \subset x$
  - Now both the hypothesis h_1 and h_2 are allowed to classify the U'
  - Add the labelled data from the previous classification to L
  - Repeat the procedures for different iterations until U is empty
- Author has used probabilistic approach that is Naive Bayes to prove the suggestion

- The Experiment has improved results by 22 percentage comparing to the existing methods, the data set is the collection of webpages from the American education institutes, where need to classify based on the departments (like computer science, mechanical etc) in addition to the faculties of the department.
- This is a True Engineering problem since the author has proved his suggestion via experiment

2 Paper-2

2.1 Reference

Wang, T.; Pouyanfar, S.; Tian, H.; Tao, Y.; Alonso, M.; Luis, S. Chen, S.
A Framework for Airfare Price Prediction: A Machine Learning Approach
2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI), 2019, 200-207

2.2 Claim

- In this article author has suggested the best algorithm to compute air fare among the various algorithms in machine learning.
- Airfare is calculated by various factors such as distance of the travel, time of the booking, fuel price, class of seat etc.Each companies calculate their own way based on the above factors to yield profit, elements related to marketing, economic and societal trends plays a vital role in deciding the airfare
- Some of the earlier methods are
  - Linear regression - which predicts with the error rate 22.9
  - SVM classification produces better result than regression
  - K-mean algorithm which clusters the similar patterns together for the price range and then CART used to extract the important features and RF arranges it to the order of purchasing
  - ANN with genetic algorithm produces a far better result with mean absolute percentage error accounts to 9.11
- Among the learning algorithm Random forest gives the better results in both evaluation metrics as 66.58
- So with the feature selection the accuracy has increased, moreover crude oil price is less sentive yet if we remove and try to predict the results then there is a drop of 7

- Even some of the macro features like transaction history, season on events or function where the demand goes high, airline pattern will help predicting the price in the daily basis instead of quarterly

3 Paper-3

3.1 Reference

Snoek, J.; Larochelle, H. Adams, R. P.
Practical bayesian optimization of machine learning algorithms
Advances in neural information processing systems, 2012, 2951-2959

3.2 Claim

- Any machine learning algorithm we use it has certain parameter such as learning rate, step size, normalization factor, hidden units etc to set for the better result of the learning algorithm. Often these are not parameter free we have to set as well as tune for better performance measure.
- In order to overcome this above problem this parameters are to be made as auto tune or optimizing the parameter on its own to satisfy this authors are suggesting Bayesian Optimization method.
- This Bayesian optimization works on the function or data that's been sampled from Gaussian process which gives the posterior distribution for the required function of our observation.
- In order to implement Bayesian optimization it faces three critical challenges those are how to make use of parallel computing, how to choose covariance function and how to reduce the computation time. The covariance function is chosen as a squared exponential kernel which is twice differential in order to smoother the output
- After choosing the covariance function the hyperparameters are taken as length scale+3
- When this algorithm is compared with TPA logistic regression using Branin-Hoo the TPA logistic regression fails is evaluating hyperparameters also in evaluating EI in time.
- SVM suffers from selecting proper covariance function when it is used along with Bayesian optimization SVM results were par from the normal SVM Running CNN on CIFAR-10 with Bayesian optimization has increased the result by 3 percentage

4 Paper-4

4.1 Reference

Coates, A.; Ng, A. Lee, H.
An analysis of single-layer networks in unsupervised feature learning
Proceedings of the fourteenth international conference on artificial intelligence and statistics, 2011, 215-223

4.2 Claim

- In current scenario, most of the work or development in Machine learning is focused on feature representation from input data which are unlabeled used for classification. So this solves multiple layers of feature, one at a time using unsupervised learning algorithm which increases the expense and complexity.
- Therefore the article suggest of using single layer with more hidden units or more features extracted from the pre-processing, this is demonstrated using four unsupervised learning methods such as Sparse auto-encoders, Sparse restricted Boltzmann machine, K-means clustering, Gaussian mixture
- Frame work for unsupervised feature learning
  - Extract few pixels of data in the random order from the image
  - Run the pre-processing on the extracted data
  - Map the feature by unsupervised learning method
  - now we have labelled data, so extract all other pixels from the input image
  - project this image in the lower dimension or particular part of the input image
  - classify the pixels based on the training
- Similarly when stride increase the performance decreases therefore the performance is inversely proportional to stride. So the author has fixed the stride as 1 ($S > 1$)
- Larger the region of image which lowers the feature extraction therefore the authors have fixed the receptive field size as 6 pixels after testing for 8 and 12 which lowered the performance.
- Based on the above parameter settings, K-means clustering has given improved results compare to other multi-layer nets. Thus the article proved simple single layer learning method with proper parameters tuned can yield better performance.

5 Paper-5

5.1 Reference

Caruana, R. Niculescu-Mizil, A.
An empirical comparison of supervised learning algorithms
Proceedings of the 23rd international conference on Machine learning, 2006, 161-168

5.2 Claim

- In this article the author has compared the performance of ten learning algorithm especially used for supervised learning, before and after calibration of parameters across eight performance metrics
- In recent times there is no empirical studies comparing modern learning algorithms like Random Forests, SVM, Boosting etc, Statlog was in early 90 after which there was no comprehensive studies to state the comparison for recent algorithms this made the authors to write this article.
- In today's scenario the Machine learning methods are used in different domains with different performance metrics which is suitable for that particular domain where it has been implemented therefore this method may give optimal performance on different performance metrics or may be not. Therefore there should be a comparison of these different methods on different performance metrics.
- The eight performance metrics are classified into three groups such as threshold metrics, ordering/ rank metrics and probability metrics based on their characteristics. Threshold metrics the performance is evaluated by certain threshold above the threshold has good performance and below are low performed. Ordering metrics orders the good performing attributes to low performing attributes. Probability metrics uses probability method therefore performance having value more or less one is good performing rest or not.
- some methods perform well on certain limited data set or certain limited performance metrics therefore this article test the scenario with bootstrap analysis in order to find the situation and results outcomes approximately the same.
- Therefore boosted trees perform well which is followed by random forest, calibrated SVM and neural nets without calibration. Naive bayes, decision tree, logistic regression are the low performer

6 Paper-6

6.1 Reference

Awad, W. ELseuofi, S.
Machine learning methods for spam e-mail classification.
International Journal of Computer Science Information Technology (IJCSIT), 2011, 3, 173-184

6.2 Claim

- Any machine algorithm tends to classify the data among a huge data set based on its learning experience if it had learned from lots of irrelevant data then the classification obviously won't output as required. Therefore filtering the relevant data as well as the feature is much import to improve the performance.
- In conceptual perspective, most facing problem is how to select the relevant feature because in the pool of relevant and irrelevant feature we need to select the most relevant features and need to combine those features together inorder to act upon the upcoming data
- Contradiction to the above point, in practical perspective, the induction algorithm (algorithm which learns based on the data everytime) converges well with more irrelevant features. Having irrelevant feature greater than the relevant feature increases the error exponentially.
- In order to select the proper feature required for the algorithm is to go with heuristic search which tends to find better features at each step of its exploration. This demands four main steps
  - Finding the starting point, this is tricky because choosing the wrong starting point may lead the convergence in the wrong direction or takes time to converge
  - Organization of search how much neighbourhood the algorithm tends to find for the given node, this size decides the time and space complexity.
  - Strategy to be followed because each approach has its own merits and demerits popularly used are greedy search and BFS former is fast and latter is a efficient
  - When to stop the search this is also a vital thing because the algorithm has to break at one point which may be achieving the required accuracy or where the accuracy haven't increased after several iterations.

- As relevant feature is important having relevant data set is also important this can be done in intensive or in conservative method. Where intensive looks for all the type of features and select the best in other hand conservative breaks once it got the best it won't look for other types
- There are now exponential time complexity in finding the relevant features which precede in the theoretical challenge, increase in efficiency even though the proportion of irrelevant feature is high and very large data set is the challenging domain to address is the imperial challenge