UNIVERSITY OF JEAN MONNET

LABORATOIRE HUBERT CURIEN

CONNECTED INTELLIGENCE TEAM

Object extraction techniques and visual image search with Semantic web techniques

Submitted by:
Aninda MAULIK,
CPS2

Supervisor:
Prof. Pierre MARET
Dennis DIEFENBACH

June 23, 2020

Abstract

This internship is about exploration of object detection and extraction techniques with a state of the art computer vision api. Thereafter, we design a semantic web model for the extracted data and finally implement a visual image search engine through Qanswer.

1. Introduction

Users' experience is an important factor for the success of a given application. Thus, the front-end of Qanswer which highly impacts the users' experience, is an important part for a image base query system. Qanswer, well handles the translation from a natural language question to correct SPARQL queries.SPARQL has emerged as the standard RDF query language. An RDF query language is able to retrieve and manipulate data stored in Resource Description Framework (RDF) format. RDF data model is based on the idea of making statements about web resources in expressions of the form subject–predicate–object, known as triples. The subject denotes the resource, and the predicate denotes traits or aspects of the resource, and expresses a relationship between the subject and the object. We generate the rdf data model from a csv file, in which each line includes information for a triplet and all its components. The csv file is generated by consolidating the information and details about the required images. We primarily get the information of the required images by running the state-of-the-art, real-time object detection system; YOLO(You Look Only Once).

2. Presentation of the research problem

There is no way to use the knowledge generated by computer vision techniques, to query image bases.The research community has made a lot of efforts to use the computer vision techniques for extracting knowledge from images. On the other side, not much attention has been paid to the implementation of methods for making this knowledge available.We hope to change this trend by presenting Semantic Web techniques for querying the knowledge made available by computer vision. My work focuses on bridging the two disciplines here.

3. State of the art

There are several research topics which either try to solve the same problem of 'small n large p ' or use the same setting or method as this one.

3.1. Variable Selection technique

This method presents an interesting way of handling the "small n, large p" problem. It employs a recursive partitioning system to determine the degree of importance of variables. The results indicate that it is not always required to correctly identify all the important variables in every situation. The algorithm is an extension of the GUIDE classification and regression tree algorithm[1, 2]

3.2. Integrating Expert Knowledge When Learning Bayesian Networks From Data

In this research topic, a method was presented to integrate expert knowledge into learning of bayesian networks. Bayesian networks typically use the dependence structure

amongst variables to graphically represent their multi-variate joint probability distribution. A Bayesian Network is a directed acyclic graph where nodes correspond to the variables in the domain problem and the edges between two variables correspond to the direct causal probabilistic dependencies. The observation model was a mutlinomial Bayesian Network with Dirichlet priors. The user input was taken as feedback to queries about the presence or absence of edges in the graph. The selection of which edge to query about next was selected by maximizing the information gain with regards to the posterior inclusion probability of the edges. Monte Carlo algorithms were used for the necessary computations [3].

3.3. Bayesian Visual Analytics

In this research topic, a framework was presented for interactive visual data exploration. Two observation models were described, principal component analysis and multidimensional scaling, that were used to reduce dimensionality for visualizing the observations on a two-dimensional plot. Their interface allowed moving points closer or further apart in a low-dimensional plot which would be encoded in a feedback model that transformed the feedback into appropriate changes in the shared parameters [4].

3.4. Bayesian nonlinear regression

The usual way to handle the 'small n large p ' problem is to reduce the number of covariates by using variable selection techniques or projecting them to lower dimensions using principal component analysis or other related methods. Most of these existing methods for variable selection or projections assume linear relationship between the response and the covariates. However this assumption is not very realistic. The model proposed here develops multivariate nonlinear regression models for accurate prediction by reproducing the kernel Hilbert spaces (RKHS) under the multivariate correlated response setup [5].

3.5. Interactive learning

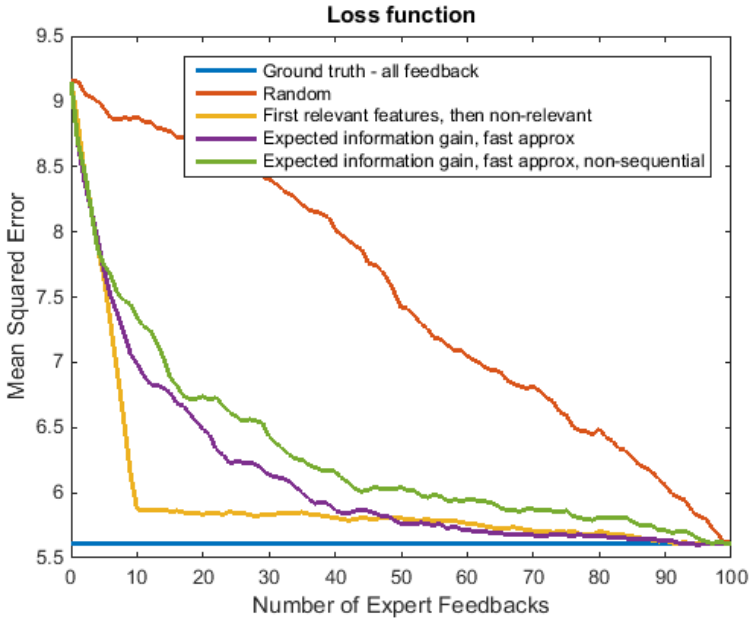
This is a model for statistical analysis where the inference procedure interacts with the data acquisition mechanism to make feedback-driven measurements. Interactive learning has been successfully applied to a variety of machine learning problems to employ user's preferences,knowledge and cognition and get better results for statistical learning. The main attraction of interactive learning is its statistical and computational efficiency. By incorporating feedback into the statistical measurement process intuitively we can understand that an interactive algorithm would achieve desired statistical performance with fewer measurements than a non-interactive one. [6]. These methods are primarily used in learning user intent and preferential clustering.

3.6. Active Learning

In active learning the learning algorithm is able to interactively query the user for the most informative data points. Usually this type of learning technique is used in unsupervised or semi supervised machine learning where data samples can be easily obtained but obtaining labels are hard. The knowledge elicitation method similarly queries the user for information trying to maximize the information gain from each feedback and thereby minimizing the required number of feedbacks required.However, the work presented here is slightly different, the goal of this research is not to gather labels for data samples but gather information about the features. Our method is more useful where obtaining additional samples is either too expensive or simple impossible.

3.7. Expert Knowledge Elicitation with Simulated Data

Synthetic Data was used to study the behavior of the proposed approach. The covariates of n training samples were generated from a normal distribution $X \sim N(0, I)$. There will m regression coefficients w_1,...,$w_m \in$R,m*=10 are generated from $w_j \sim N(0, \psi)$ and the rest are set to zero. The observed output values were generated from the normal distribution $y \sim N(Xw, \sigma^2 I)$. Two cases were considered, one where the user has knowledge about relevant/non-relevant features with γ_j=1 if w_j is non-zero and γ_j=0 otherwise, π=0.95. The other case was where the user has knowledge about the value of the coefficients(with noise w=0.1). The mean square error(MSE)has been calculated by taking the mean of the square of the differences between true values of the output and its emperical values [7].



The four query algorithms have been compared

1. random feature suggestion
2. an oracle(unrealistic) strategy that knows relevant features from beforehand and asks exclusively about the first and then at random chooses features that have not been already selected
3. the sequential experimental design
4. a non-sequential setup of the feedback model

The use of additional knowledge in the form of expert feedback certainly reduced the prediction error from the very first use of feedback and it performs significantly better than random feature selection and the non-sequential setup.

4. Proposed Model

This internship is an extension of the research topic 'Knowledge Elicitation via the Sequential Probabilistic Inference for High-Dimension Prediction' [7]. The existing model takes input from the expert who either has knowledge of the coefficients of the covariates or the relevance of the coefficients in the form of not-relevant, relevant or uncertain features. This input is incorporated in the feedback model and is used to sequentially update

the posterior distribution. The model proposed here considers the pool of experts to be voters in a majority vote on each feature of the dataset.This majority voted knowledge is then incorporated in the form of feedback and is used to sequentially update the posterior distribution as before.

Algorithm 1 Knowledge Elicitation

Input: $D = (y_i, x_i) : i = 1, .., n$
Output: $p(f_{t+1}|D, f_1, .., f_t)$
1: Calculate $p(\theta|D)$ using
2: **loop** $(f = 1 : t)$
3: Take feedback from the experts and do a majority vote on each feature for its inclusion or exclusion
4: Sequentially update the posterior distribution to obtain $p(f_{t+1}|D, f_1, .., f_t)$
5: **end loop**

5. Experiments

The proposed model was tested on a dataset with Amazon products in the kitchen appliances category.

5.1. Dataset used

Amazon Data The Amazon data is a subset of the sentiment dataset of [8]. The dataset contains textual reviews and corresponding ratings 1-5 for Amazon products. Here, we have only considered reviews for products in the kitchen appliances category amounting to 5149 reviews. Each review is represented as a vector of features, for each distinct feature and each review, a matrix of occurrences was created and only those features were kept which appeared in atleast 100 reviews, a total of 824 features. For user study, ten university students and researchers were asked feedback on all the 824 features in the form of not-relevant, relevant or uncertain. It was assumed that the algorithm had access to 100 training data and at each iteration it could query the pre-given feedback one word at a time. The dataset was partitioned into three parts:

1. a training set of 100 randomly selected reviews
2. a test set of 1000 randomly selected reviews
3. a "user-data-set" for constructing simulated user-knowledge [7].

5.1.1. Preprocessing of the Amazon dataset

The same pre processing techniques that were used by [7] were used here. At first the review file was read, the bag of words of the full data were prepared and the response values were extracted. Then,keywords with less than 100 occurrences in the reviews were removed. Finally, cross validation on the data was done to learn the best parameters for Spike-and-Slab linear regression. After that, the best parameters were used to fit a model to the data. The output was used as some kind of a ground truth of data.

5.2. Experimental results

The following are the mean squared errors for:

1. 10 individual real experts
2. 5 majority voted experts
3. 10 majority voted experts

The model parameters were set to π=0.7,ψ^2=0.01, $\alpha_\sigma = 1$,$\beta_\sigma = 1$, $\rho = 0.3$ (Eq. 4)

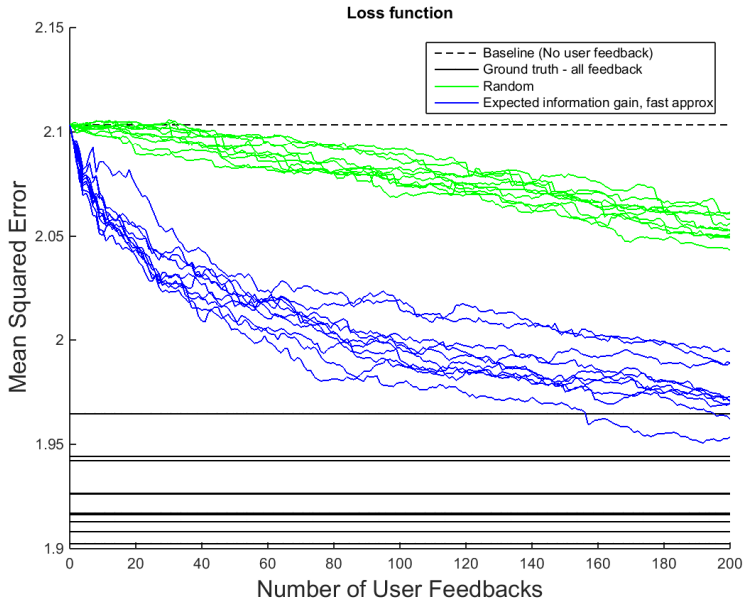


Figure 1. Mean squared error for 10 individual experts

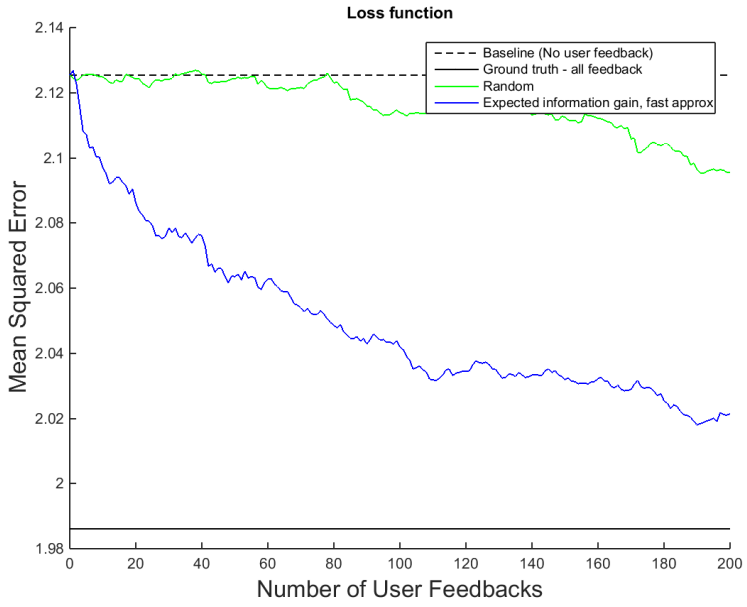


Figure 2. Mean squared error for 5 majority voted experts

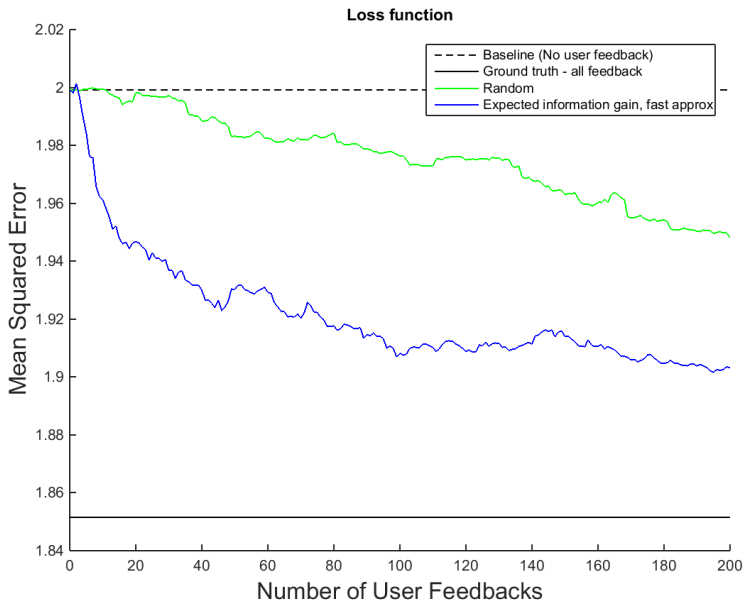


Figure 3. Mean squared error for 7 majority voted experts

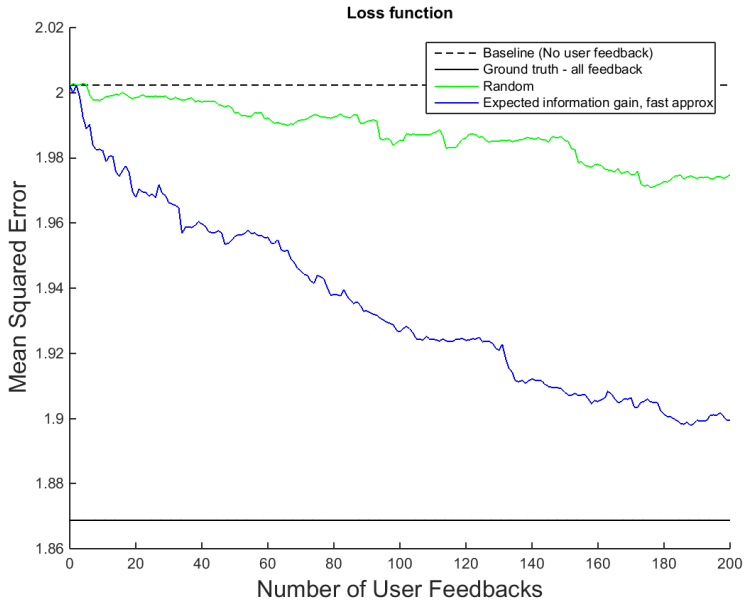


Figure 4. Mean squared error for 10 majority voted experts

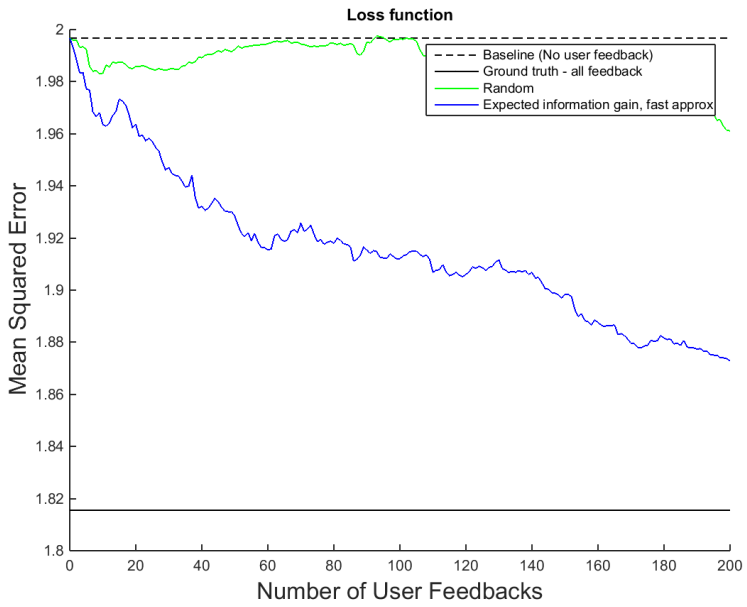


Figure 5. Mean squared error for 10 majority voted experts with π=0.8

It is clear that the use of expert knowledge via the sequential query model improves the average mean squared error at a better rate than random feature suggestion. The sequential model shows a faster rate of improvement than the random strategy which suggests that it asks about the most important features first.
While comparing the 5 majority voted, 7 majority voted and 10 majority experts, we see that for 5 majority voted experts the results were similar to individual experts being used for prediction, there was no real improvement in the mean square error values. However, with increase in the number of majority voted experts to 7, there was a slight improvement in the MSE values and for 10 majority voted experts the improvement was quite visible.Also, there was also a sharper improvement in the MSE values with increase in the number of feedbacks from the very beginning for the 10 majority voted experts case. This seems to suggest that with increase in the number of experts, a majority vote strategy on the features seems to bring about better results. Also, with increase in the confidence coefficient of the experts there seems to be a slight improvement in the mean squared error.

6. Conclusion

A majority voted knowledge elicitation problem was presented for solving the ”small n, large p” problem in a sparse linear regression model. The model showed improved prediction accuracy(MSE reduction) even with just a few feedbacks when compared to the model with random feature selection. It also performed slightly better with increase in the number of experts for majority vote. Also, the prediction error reduced slightly when the confidence coefficient was increased. This method reduces the individual biases of the experts for better prediction accuracy. The presented method is generic and all the assumptions are probabilistic, thus it can be be tailored to suit other elicitation settings.

7. Future Works

A more extensive research can be done with more number of experts with different levels of expertise to study the behavior in the reduction of the mean squared error.

References

[1] Wei-Yin Loh. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14–23, 2011.

[2] Wei-Yin Loh. *Variable Selection for Classification and Regression in Large p, Small n Problems*, pages 135–159. Springer New York, New York, NY, 2012.

[3] A. Cano, A. R. Masegosa, and S. Moral. A method for integrating expert knowledge when learning bayesian networks from data. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(5):1382–1394, Oct 2011.

[4] Leanna House, Scotland Leman, and Chao Han. Bayesian visual analytics: Bava. *Statistical Analysis and Data Mining*, 8(1):1–13, 2015.

[5] Sounak Chakraborty, Malay Ghosh, and Bani K. Mallick. Bayesian nonlinear regression for large p small n problems. *Journal of Multivariate Analysis*, 108(C):28–40, 2012.

[6] Joseph B. Kadane, James M. Dickey, Robert L. Winkler, Wayne S. Smith, and Stephen C. Peters. Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association*, 75(372):845–854, 1980.

[7] Pedram Daee, Tomi Peltola, Marta Soare, and Samuel Kaski. Knowledge elicitation via sequential probabilistic inference for high-dimensional prediction. *CoRR*, abs/1612.03328, 2016.

[8] John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In *In ACL*, pages 187–205, 2007.