VOL. 13, NO. 20, OCTOBER 2018 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

CONDITIONAL RANDOM FIELDS AND REGULARIZATION FOR EFFICIENT LABEL PREDICTION

Richa Chaturvedi, Deepak Arora and Pawan Singh
Computer Science and Engineering Department Amity School of Engineering Technology, Amity University, Uttar Pradesh, Lucknow, India
E-Mail: richachaturvedi14@hotmail.com

ABSTRACT

Natural language processing task usually involves predicting a large number of variables that depend on each other as well as on other observed variables. We have studied different approaches: generative and discriminative that can be taken into consideration. CRFS, HMMs and MaxEnt can be used but, CRFs particularly have seen wide application in this particular area. CRFs can also be used in computer vision, and bioinformatics. Moreover, regularization plays a vital role and L1 and L2 regularizes are critical tools in machine learning due to their ability to simplify solutions. In this paper we compare both L1 and L2 regularization technique while they are being applied on dataset consisting of news articles. The results obtained are checked by three parameters namely Precision, recall and support.

Keywords: CRF, HMM, MaxEnt, regularization, F measure, precision, recall.

1. INTRODUCTION

Semi-supervised learning algorithms are trained using both unlabeled and labeled data. The presence of data which is unlabeled is more in comparison to data which is labeled. Thus, Semi-supervised learning is said to be falls between unsupervised learning and supervised learning. This is the main reason why semi supervised algorithms are best suited for problems and scenarios related to NLP. In scenarios of NLP often there is a plethora of unlabeled data and only a smaller quantity of labeled data is actually present for the training purposes also, the whole process of labeling the unlabeled data can be costly and tedious as well. And labeling this data can sometimes result into adding some human biases. So if there will be a massive unlabeled data during the training process it will be beneficial for the final model as the veracity of the model increases while decreasing the time and effort put into it. Semi-supervised learning is more preferred in cases like webpage classification, sentiment analysis and genetic sequencing. In all these cases gathering massive amount of unlabeled data of {x} is actually very easy. On the other hand, collecting their corresponding labels {y} for making predictions is slow and time consuming [1]. This labeling bottleneck results in the presence labeled data in a smaller amount and an excess of unlabeled data. Therefore, it will be beneficiary for us to utilize excess amount unlabeled data is [Xiaojin Zhu].

In the past, semi supervised learning can be conducted using various different models which include: generative models, cotraining [11] and graph based transductive methods [12]. However, all these proposed approaches are suitable only for single class label classification problems input [12].

2. LITERATURE REVIEW

Classification algorithms typically are of two types: Generative models or Discriminative models. In probability theory for the calculation of probability of occurrence of an event A given B (i.e. P (A|B)), it calculates the likelihood of event B given A (i.e. P (B|A)) and probability of event A and using Bayes theorem (i.e. PA|B) = P (B|A) * (P (A) / P (B)) finds required probability. We can hereby note that, a generative algorithm has to learn how the data is generated in order to categories a data point. This is like finding an answer to a more general problem as a transitional step towards finding the solution of the target problem. Whereas, a discriminative model takes an easy and straightforward approach towards solving the target problem, all it requires is quality data and makes almost no assumption on the distribution of data.

In a case where there is plethora of clean data, a discriminative model will always outperform a generative model [Andrew Y.Ng and Michael I. Jordan]. There are many imperative reasons on using discriminative model [2] over generative model firstly, generative models usually cannot achieve the same accuracy as achieved by discriminative models, and secondly, as stated by Vapnik [3] we should approach towards taking care of the issue specifically and never pick to take care of a general issue an intermediate step. Of course, we here are not taking into consideration the computational issues that can arise or problems like taking care of the missing data.

Both HMM, which is Generative, and CRF, which is Discriminative can be used in classification of data. But when the data is massive and enormously large Conditional Random Field has always outperformed the HMM and MaxEnt [Manish *et al.*]. The pretext behind this huge difference in performance is that CRF is a sequence classifier for structural prediction. Input for CRF is a sequence of phrases a1, a2,.. am. Here we have to predict the tag b1, b2 . . . .bm for each word. Like HMM, it also predicts the most probable sequence of tags using Viterbi Algorithm and then trains the model using gradient descent but imperative difference is that in HMM current word tag only depends on the word before the current word tag, and the current word is dependent on the current tag only. HMMs presume that each word does not depend on its context which obviously is contrary to the fact. In

spite of these drawbacks, the HMM model will still give a number of advantages such easiness and quick learning [5]. This tells us, here we can only make us of a small set of local features, and on the other hand there are no such constraints for CRF, which makes it powerful from the two of the major statistical techniques applied. CRF outperforms HMM and HMM outperforms MaxEnt, which was attributed because of the presence of the label bias problem [Lafferty *et al*].

MaxEnt are often preferred in the cases where we want to feed some kind of auxiliary information to the model but MaxEnt have a limitation that makes it more irresolute than HMM when no additional information is used. For example statistical model shown in Figure-1. If we take into account all the local probabilities for each observation we can deduce that the transition towards state 2 has always the highest probability for state 1. We can see that the same happens with state 2. However, if in the same scenario if make use of the Viterbi algorithm in order to get to the path that is most probable we come out that it is the path 1 → 1 → 1 → 1. This scenario occurs because here we have only two transitions out of state 1 compared to five transitions out of state 2. The probabilities of the transition from a given state in MaxEnt are normalized therefore; the model will prefer the states with a lower number of transitions and this will ultimately give an unfair advantage to some particular states and result into a bias towards the states that have lesser transitions that are going out. In severe situations, state that has a single transition that is going out can completely breach the observation. This phenomenon is called the label bias problem.

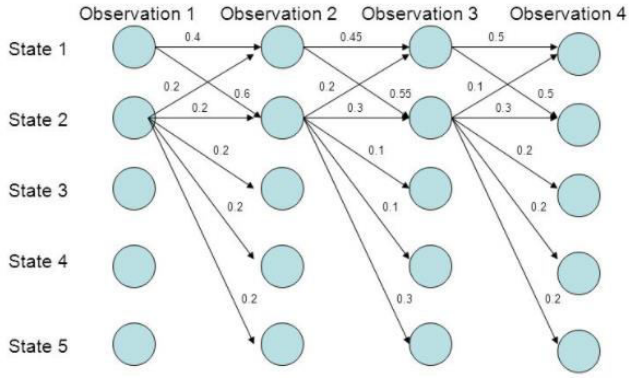

Figure-1. A statistical model.

CRF thus offers all the positives of MaxEnt while overcoming the label bias problem. The difference between CRF and MaxEnt is that the MaxEnt model makes use of conditional probabilities exponential model for each state. With this model it is possible to normalize the probabilities at a global level. CRF can be said to have the strengths of HMM and MaxEnt in the sense that complex features of MaxEnt can be highly correlated are incorporated and there is no assumption of independence as well state transition features of HMMs.

CRFs are thus very useful in scenarios of labeling tasks, because as conditional models, they don't take into considerations that are typically made by traditional generative models like MaxEnt and HMMs. Conditional Random Fields tend to be advantageous over others because it gives flexibility for using overlapping features and at the same time giving two advantages: first, can be used in both classification and training, and secondly permitting the global optimization. WE tend to use CRFs more in scenarios of natural language processing because they can easily integrate arbitrary and features that are not so independent but are still present in the input without making any assumptions among the features. This is very important particularly for sentiment analysis. The approach is checked using these three criteria: precision, recall, precision and F-measure.

Conditional random fields (CRF) is particularly defined as an undirected graphical model which can be represented by (A, B, G, g, θ). Where A = (A1, A2...) is a set of input variables and B = (B1, B2...) is a set of output variables. We use a = (a1, a2...), to denote a possible assignment of values to A, and similarly for b, and ab denotes the joint assignment. For every $\delta \in$ G, the CRF gives a specific function g that withdraws a feature vector $\in R^d$ from the restricted assignment ab. The model defines conditional probabilities

$$p_\theta(b|a) \frac{= \exp \vec{\theta} \cdot \vec{g}(a,b)}{\sum_{b\prime} \exp \vec{\theta} \cdot \vec{f}(a,b\prime)} \quad (1)$$

Where, $\theta \in R^d$ denotes global weight vector that has to be learned. The denominator sums up to every possible output assignment to normalize the distribution, as this belongs to log linear model.

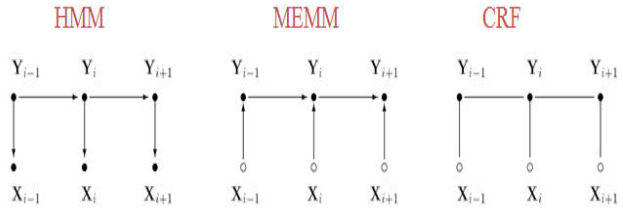

Figure-2. Graphical structure of HMM (left), MaxEnt (center) and Chained CRF (right) for sequence.

Figure-2 shows graphical structure of HMM, MaxEnt or MEMM and chained CRF. Y_0 denotes start states and Y_{n+1} denotes stop state.

2.1 Over-fitting and regularization

In semi-supervised machine learning algorithms, models are trained on training data. The comprehensive objectives of the process are to find the target of each training example from the training data. Overfitting often happens and we try to regulate or decrease it. Overfitting occurs when a model learns the noise as well as the signal in the training data and thus in result it will not be able to perform that well on the new data just given on which model wasn't trained on. We can avoid overfitting in the model on training data by either reducing number of features cross-validation or regularization.

Methods like cross-validation and regression will do feature selection which will work well with feature where there is comparatively smaller set of features available but techniques like regularization comes handy when the scenarios consist of large set of features. Regularization will basically add some penalty as model complexity increases. Our final goal is to create a less complex when we are having a large number of features in our dataset.

The Regularization techniques we are going to discuss in this paper are: L1 Regularization and L2 Regularization. The vital difference between these two techniques is the penalty term and that L1 diminishes some features and ultimately nullify them thereby eliminating some features altogether. So, this works well for feature selection in case we have a huge number of features.

2.2 L2 regularization

A regularization model that makes use of L2 regularization is called Ridge Regression. Ridge regression will add penalty expression to loss function. L2 Regularization is used to penalize CRFs while training the model. Here instead of only maximizing the conditional likelihood we subtract the penalty expression for every weight that is directly proportional to w_2^i from the likelihood:

$$max\, l(\ B|\, A;\ w)\ -\ \lambda wT\ w \qquad (2)$$

The term λ here will control the smoothness that is being applied to the model. A higher value of λ means smoothing will be more and value of λ zero will results into no smoothing at all. The only difficulty arises while selecting an accurate value for λ; this accurate value is achieved by checking for different cases of λ by cross-validation. Here the penalized objective function will remain to be differentiable and thus training a CRF with an L2 regularization will require equivalent computational efforts compared to training a CRF without regularization.

2.3 L1 regularization

A regularization model that uses L1 regularization is called Least Absolute Shrinkage and Selection Operator (Lasso) and it wholes up to be the coefficients absolute value as penalty term to the loss function.

Instead of penalizing the objective function byw_2^i, as performed by the L2 regularization, here the L1 regularization will take into account the penalties that are proportional to
|wi|, which will result into the penalized objective function:

$$\max l(\ \mathrm{B}|\ \mathrm{A};\ \mathrm{w}) - \lambda\textstyle\sum|\mathrm{w_i}| \qquad (3)$$

Here λ again denotes a term that will limit the degree of smoothness applied during training steps. Here the penalized objective function is no longer differentiable. L1 penalty regularization will complicate the training, but still it does have certain profit of producing sparse models [8]. This infers to the fact that L1 regularization sometimes result into making models where some weights become nullified which is somewhat similar to feature selection because features where weights are nullified tend to have no effect on classification and thus can be easily eliminated from the model.

Recent work in this area has proven that L1 regularization is usable in generative models, the group of models that incorporates CRFs. Douglas L. Vail and John D. Lafferty have concluded that [9] L1 regularizers is an powerful method for feature selection in CRFs. Feature selection plays an crucial role in cases where classifiers are hugely dependent on complex features.

2.4 L1 vs. L2 regularization

Though L1 and L2 regularization are hugely similar and there are some theory guidelines about which form of regularization to choose in certain cases, the bottom line is that one must experiment and find which type of regularization suites his case better and moreover, whether making use regularization affects the overall result. Though L1 and L2 regularization are hugely similar and there are some theory guidelines about which form of regularization to choose in certain cases, the bottom line is that one must experiment and find which type of regularization suites his case better and moreover, whether making use regularization affects the overall result. The key distinction between the L1 and L2 is only that L2 is the entirety of the square of the weights, while L1 is only the aggregate of the weights.

Although using L1 regularization technique can have a response which can be beneficial in making one or more weight values to 0, which will mean that the associated feature isn't needed anymore. This is one form of what's called feature selection. Whereas, L2 regularization will limit model weight values, but usually doesn't prune any weights entirely by setting them to 0.0 thus not giving any implicit feature selection.

Thereby it may seem to us that L1 regularization is better than L2 regularization. But there still is a disadvantage of L1 regularization that the technique can't be easily used with some ML training algorithms, particularly algorithms that make use of calculus in order to compute gradient on the other hand, L2 regularization can be used with any type of training algorithm.

3. EXPERIMENT

While checking the effectiveness of L1 and L2 regularization on CRF model for label prediction, we performed sequence labelling in NLP which aims at classifying the correct label for each named entity present in the dataset. The dataset used was 500 news gold standards; it consists of 500 English news articles from online news platforms. This dataset is already labeled, but many labels are missing. There are two labels N and I, N stand for part named entity and I stand for irrelevant word respectively. Then we trained the CRF model using the part-of-speech (POS), for this we made us of the NLTK'S

POS tagger. Python CRF suite was used in training the CRF classifier. After the POS tagging, we can generate more features for the tokens in the dataset. The result is checked on the basis of precision, recall, f1 score and support where,

Precision= $\frac{tp}{tp+fp}$,

Recall= $\frac{tp}{tp+fn}$,

F measure= 2. $\frac{precision.recall}{precision+recall}$

Here, tp term is denoting the value for true positive, fp denotes the value for false positive and fn denotes values for false negative.

4. RESULT

We take into considerations four parameters namely precision, recall, f1 score and support. When we change the value of coefficients L1 and L2 some changes are seen, while these changes are small but can be quite effective. In Figure-3 we can see the values of precision, recall, f1 score and support for the first case when we applied L1 regularization. Figure-4 displays the values for precision, recall, f1 score and support for the second case when we applied L2 regularization. In the first case we can see that we have achieved 61% precision and 20% recall in predicting whether a word is part of a named entity while in the second case we can see that we have achieved 69% precision and 68% recall in predicting whether a word is part of a named entity. Clearly, for our dataset L2 regularization gives much better results compared to L1 regularization. Figure-5 and Figure-6 is showing the graph between precision and recall in each of the cases.

```
as (I)
a (I)
new (I)
federal (I)
program (I)
began (I)
. (I)
              precision    recall  f1-score   support

          I       0.90      0.98      0.94      2758
          N       0.61      0.20      0.30       394

avg / total       0.86      0.88      0.86      3152
```

Figure-3. Values for precision, recall f1-score and support with L1 regularization.

```
appalachian (I)
power (I)
customers (I)
in (I)
west (I)
virginia (I)
. (I)
              precision    recall  f1-score   support

          I       0.95      0.95      0.95      2478
          N       0.69      0.68      0.69       370

avg / total       0.92      0.92      0.92      2848
```

Figure-4. Values for precision, recall f1-score and support with L2 regularization.

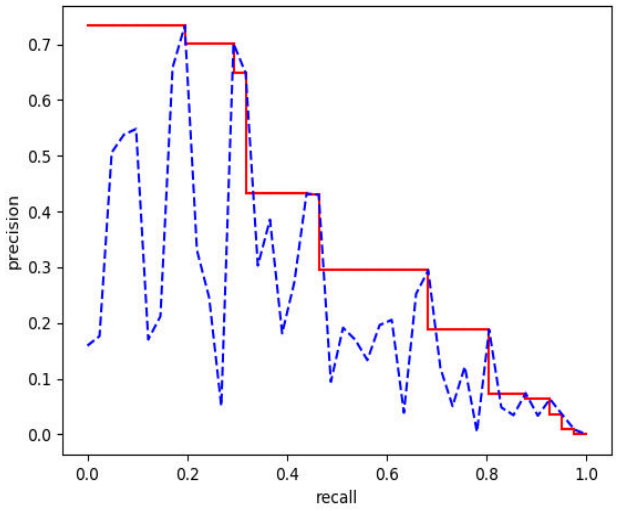

Figure-5. Graph between precision and recall for L1 regularization.

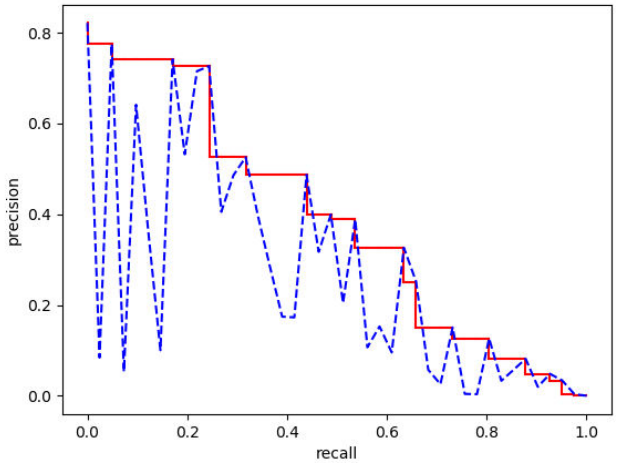

Figure-6. Graph between precision and recall for L2 regularization.

5. CONCLUSION AND FUTURE WORK

Authors have compared and studied the Conditional Random Field or CRF and compared CRF to HMM and MaxEnt and found out the reasons which supports that CRF are indeed best to be used in scenarios where natural language processing is involved. We have

further studied the regularizations techniques that can be done on CRF and found out that regularization can be used with any ML classification technique that's based on a mathematical equation and that regularization eliminates the magnitudes of the weight values in a model and thus sometimes called weight decay. The major role of using regularization technique is that it results into more accurate model as it eliminates over fitting. We further studied the two type of regularization L1 and L2 and compared them by applying on the data. In our case, L2 regularization seems give us better result. Although, there are some theory guidelines about which form of regularization to choose in certain cases, the bottom line is that one must experiment and find which type of regularization suites his case better. Moreover, there are several things by which we can improve the performance, including creating better features or tuning the limitations of the CRF models. We can also try to incorporate the numerical features like, number of characters in the model.

REFERENCES

[1] Zhu Xiaojin & B. Goldberg. 2009. Introduction to Semi-Supervised Learning. Morgan & Claypool Publisher, ISSN 1939-4608

[2] Ng Andrew Y. and Jordan Michael I. 2001. On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes, Advances in Neural Information Processing Systems 14, Massachusetts Institute of Technology Press. 841-848.

[3] Vladimir N. Vapnik. Sep 1998. Statistical Learning Theory. John Wiley & Sons Inc, NY.

[4] Agarwal Manish, Goutam, Rahul, Jain Ashish, Reddy Kesidi, Sruthilaya, Kosaraju, Prudhvi, Muktyar Shashikant, Ambati Bharat Ram & Sangal Rajeev. 2018. Comparative Analysis of the Performance of CRF, HMM and MaxEnt for Part-of-Speech Tagging, Chunking and Named Entity Recognition for a morphologically rich language. Language Technologies Research Centre International Institute of Information Technology.

[5] John Lafferty, Andrew McCallum, and Fernando C.N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proc. of the 18th International Conference on Machine Learning, pp. 282-289, MA, USA.

[6] Stoyanov Veselin & Eisner Jason. 2012. Minimum-Risk Training of Approximate CRF-Based NLP Systems in Proc. Of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Montreal, Canada. pp. 120-130.

[7] Simões Gonçalo, Galhardas Helena & Coheur Luisa. 2018. Information Extraction tasks: a survey.

[8] Trevor Hastie, Robert Tibshirani and Jerome Friedman Hastie. 2009. The elements of statistical learning: data mining, inference, and prediction, Springer Series in Statistics New York.

[9] D.L. Vail, M.M. Veloso and J.D. Lafferty. 2007. Conditional random fields for activity recognition. In Proc. of the 6th international joint conference on AAMAS '07 Article No. 235, ACM, New York, United States of America.

[10] Kwangmoo Koh, Seung-Jean Kim and Stephen Boyd. 2007. An interior-point method for large scale ℓ1-regularized logistic regression. Proc. in Journal of Machine Learning Research. 8: 1519-1555.

[11] Blum Avrim and Tom M. Mitchell. 1998. Combining Labeled and Unlabeled Data with Co-Training. In Proc. of the 11th annual conference on the eleventh annual conference on Computational learning theory, ACM, New York, United States of America. 92-100.

[12] Dengyong, Zhou and Schölkopf, Bernhard. 2004. Learning from Labeled and Unlabeled Data Using Random Walks. Proc. In: Pattern Recognition 26th DAGM Symposium, Tübingen, Germany

[13] T.Morgan *et al.* 2009. Data mining: Concept and Techniques.