News Article Classification

Lauren Contard, Archit Datar
Yue Li, Robert Lumpkin, Haihang Wu

1 Introduction and Problem Statement

2 Methods

2.1 KNN Methods – Binary Relevance and ML-KNN

Solutions to the general multi-label learning problem include a wide variety of approaches. Arguably, the most intuitive among these is what's referred to as "binary relevance". This approach works by decomposing the multi-label learning task into a number of independent binary learning tasks (one per class label) [8]. Binary Relevance methods are often criticized in the literature because of their label independence assumption, producing a potential weakness of ignoring correlations among labels [5]. In this project, we compare a binary relevance KNN model to a novel, bayesian KNN based approach: ML-kNN.

The ML-kNN (Multi-label k nearest neighbors) model is derived from the traditional k nearest neighbors (kNN), except for the multi-label case. While the goal of the traditional kNN algorithm is to predict whether class of the test sample based on the classes of its k nearest neighbors, the goal of ML-kNN is to predict multiple classes based on the classes of the k nearest neighbors of the test point. For the unseen data point, its nearest neighbors are identified. Then, based on the number of neighboring instances belonging to each possible class, maximum a posteriori (MAP) principle is utilized to determine the label set for the unseen instance.

ML-kNN is used in a variety of problems, such as, text categorization [7], where each document may belong to several topics, such as the use case for our project. Apart from this, it can also be useful in areas such as functional genomics where each gene may be associated with a set of functional classes [2], and in image classification, where each image could have multiple genres.[1]

The basic concept of ML-kNN is to estimate the probability that a test instance (t) has a label (ℓ) given the number of nearest neighbors ($\vec{C}_t(\ell)$ out of $N(x)$) that have label ℓ. Let E_j^ℓ ($j \in \{1, ..., K\}$) denote the event that, among the K nearest neighbors of t, there are exactly j instances which have label ℓ. Further, let H_0^ℓ denote the event that test instance t does not have a label ℓ and let H_1^ℓ denote the event that it does have label ℓ. Further, let $\vec{y}_t$ be the vector containing the predicted labels at test instance t. Then, the predicted value for the ℓ^{th} label is given as:

$$\vec{y}_t(\ell) = \operatorname*{argmax}_{b \in \{0,1\}} P\left(\mathrm{H}_b^\ell | E^\ell_{\vec{C}_t(\ell)}\right), \quad \ell \in \mathcal{Y} \tag{1}$$

Using Bayes theorem, this can be written as:

$$\vec{y}_t(\ell) = \operatorname*{argmax}_{b \in \{0,1\}} \frac{P\left(\mathrm{H}_b^\ell\right) \cdot P\left(E^\ell_{\vec{C}_{t(\ell)}} | \mathrm{H}_b^\ell\right)}{P\left(E^\ell_{\vec{C}_{t(\ell)}}\right)} \tag{2}$$

Since $E^\ell_{\vec{C}_t(\ell)}$ is independent of H_b^ℓ, this can be equivalently written as:

$$\vec{y}_t(\ell) = \operatorname*{argmax}_{b \in \{0,1\}} P\left(\mathrm{H}_b^\ell\right) \cdot P\left(E^\ell_{\vec{C}_{t(\ell)}} | \mathrm{H}_b^\ell\right) \tag{3}$$

The details of the derivation of $P\left(\mathrm{H}_b^\ell\right)$ and $P\left(E^\ell_{\vec{C}_{t(\ell)}} | \mathrm{H}_b^\ell\right)$ are beyond the scope of this report and are provided in [11].

2.2 Linear Dimension Reduction (PCA)

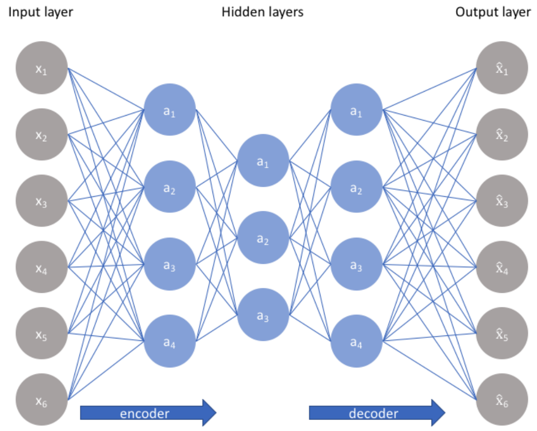
The available data for our project is high dimensional; it has a few hundred data points and a few thousand features. Such data can have significant generic disadvantages such as being pront to overfitting. More specifically, for a kNN-type model, this can cause problems because the distance the between points is considered Euclidian. In such a scenario, having an extremely high dimensional feature space causes the distances between points to be fairly similar, as these distance vector components are partitioned across many dimensions. [9]

Thus, dimensionality reduction was deemed beneficial. We performed the traditional principal components analysis (PCA) to understand if the data could be represented in fewer linear combinations. The standard PCA method partitions the total variance of the data along various linear combinations of the features.

2.3 Nonlinear Dimension Reduction (ANN Autoencoder)

In addition to reductions in dimension due to PCA, we also implement an ANN autoencoder (see section 2.4 for an introduction to ANNs). Autoencoders can learn data projections with suitable dimensionality and sparsity limitations that are more useful than other fundamental methods such as PCA, which only allow for linear data representations [6].

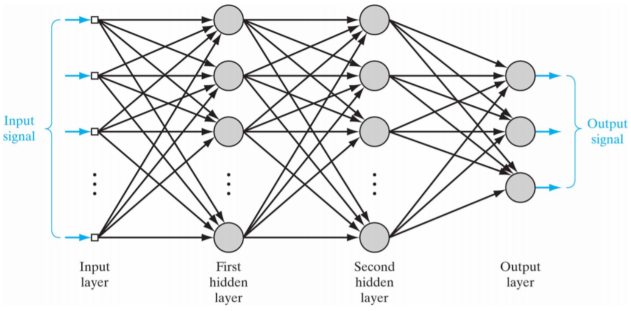
This nonlinear dimension reduction is done by by training a feed forward (FF) neural network to perform the identity mapping, where the network inputs are reproduced at the output layer. The network contains an internal "bottleneck" layer (containing fewer nodes than input or output layers), which forces the network to develop a compact representation of the input data, and two additional hidden layers [4]. Look to the diagram to the right, for a visualisation of this set-up.



The particular network that we trained had three hidden layers, as in the diagram. The first, second and third hidden layers are of dimensions 128, 64, and 128, and use the activations tanh, ReLu, and sigmoid, respectively. Training was performed using Adam optimization, MSE loss, and over 400 epochs. After training, the generated encodings were used to repeat our model fitting procedures for both the binary-relevance KNN and ML-KNN algorithms.

2.4 Artificial Neural Networks (Feed-Forward & Recurrent)

Classification approaches utilizing different artificial neural networks are also utilized in our project. Inspired by biological nervous systems, neural networks date back to the first half of the 20^{th} century with works such as those by McCulloch and Pitts, which could model simple logical operations [10]. Since most subsequent work in the following two decades centered around single layer networks, the power of neural networks was restricted to linearly separable problems. This excluded the possibility of learning even simple functions like XOR, which required a second layer [3]. In the early 1980s, research on neural networks resurged largely due to successful learning algorithms for multi-layer neural networks and are used today for various tasks such as computer vision, associative memory, representation learning, NLP, etc..



For our project, we tried implementing both FF networks and recurrent neural network (RNN) architectures. Perceptrons are the building block for FF networks. A perceptron has an input layer of neurons and an output neuron. Weighted connections exist between every input neuron and the output neuron. A forward pass through the network consists of multiplying the values of the input neurons by the value of their connections' weight, summing and then applying some non-linearity (which we call the activation function). Feed forward networks are typically built by stacking perceptron units vertically and horizontally. All of our FF networks utilize the same architectures; the first hidden layer is of size 32 with ReLU activation, dropout regularization with drop probability of 0.5, and an output layer of size 13 (the number of labels) with sigmoid activations. The FF architectures were trained on 2094 dimensional tf-idf vectors.

While FF networks display incredible expressive power, RNNs are a popular adaptation for NLP problems because they are uniquely suited for sequence processing. This is due to their hidden unit connections with shared weights, which allow for information from previous and/or future states to influence the current state. This nice feature also leads to the exacerbation of the "gradient exploding/-vanishing" problem during training. Many methods have been proposed to overcome this, one of which is the "gated" architecture we use, known as the Long-Short-Term-Memory (LSTM) architecture.

On the reduced dataset, all of our RNN models utilize the same bidirectional LSTM architecture. On the full dataset, all of our models utilize the same single-directional LSTM architecture. In both the bidirectional and single-directional LSTM's, we used hidden states of size 16 followed by a dense output layer with sigmoid activations.The RNNs were trained on padded sequences of integers corresponding to sequences of the processed tokens from the paragraphs; urls, punctuation, and stop words were all removed.

All of our networks (both FF and RNN) utilize the same optimization algorithm (Adam optimization – a variant of gradient descent), however, training was performed, using two different loss functions. The standard binary cross entropy loss function was used, in addition, to the more novel BPMLL (back prop for multilabel learning) loss. The BPMLL loss function requires instances to have at least one label. Thus, on the one hand, we trained both cross entropy and BPMLL models on such a reduced dataset in addition to training just cross entropy models on the full dataset.

The BPMLL loss function aims to leverage correlations between labels by evaluating the error as a function of pairwise errors between labels. Namely, the loss function is given by:

$$E = \sum_{i=1}^{m} E_i = \sum_{i=1}^{m} \frac{1}{|Y_i||\overline{Y}_i|} \sum_{(k,l)\in Y_i \times \overline{Y}_i} \exp(-(c_k^i - c_l^i))$$

where $c_j^i = c_j(x_i)$ is the output of the network on x_i on the j^{th} class. The back propagation algorithm is derived in the same manner is for a cross entropy or MSE loss. Details are ommitted here, but can be found in [12].

2.5 Threshold Function Learning

In both of their papers, introducing the BPMLL and ML-KNN algorithms, Zhang & Zhou also describe a common method for learning threshold functions. Perfect classification, using a constant threshold requires two conditions: (1) Logit values corresponding to labels included in an instance's label set be separated from logit values corresponding to labels not in an instance's label set. And (2) this separation be around some constant value (usually either 0.5 or 0). Learning a threshold function aims to relax the second condition. Namely, we fit a linear regression model to learn threshold values from the logit outputs of our models. For more details, see [12] & [11].

3 Results

3.1 ML-KNN Results

3.2 Artificial Neural Network Results

As some instances in the full dataset lack labels, the 'Reduced Dataset' is created by removing these instances without any labels. The full dataset and reduced dataset are then used to train and validate four networks: Feed Forward network with cross entropy loss(CE FF), Feed Forward network with BPMLL(BPMLL FF),Recurrent neural network with Bidirectional (single directional on full dataset) LSTM and cross entropy loss(CE RNN), and Recurrent neural network with Bidirectional (single directional on full dataset) LSTM and BPMLL(BPMLL RNN). We also train the neural network using 3 different training rate: 0.01,0.001,0.0001. For each learning rate, the validation set hamming loss is plotted against epochs as shown in figure 1 and 2.

Figure 1 shows the training history of CE FF and CE RNN using full dataset and constant threshold. From the left and middle sub figures, it can be shown that CE FF can achieve lower hamming loss than CE RNN. For the right sub figure, though the CE FF has higher hamming loss than CE RNN, CE FF seems to have decreasing trend of hamming loss as opposed to the plateau trend of CE RNN at 100 epoch, meaning that it is highly possible that CE FF will have lower hamming loss than CE RNN if the training continues. The effect of learning rate on training is also clear from Figure 1: smaller learning rate leads to slower training process. But from figure 1, it is unclear whether smaller learning rate can reduce the hamming loss finally.

To compare with constant threshold, table 1 gives the validation set hamming loss after 100 epochs of training, using a learned threshold function, for each network with 3 different learning rate. No significant improvement in hamming loss has been observed for both 2 networks after learned threshold function is used.

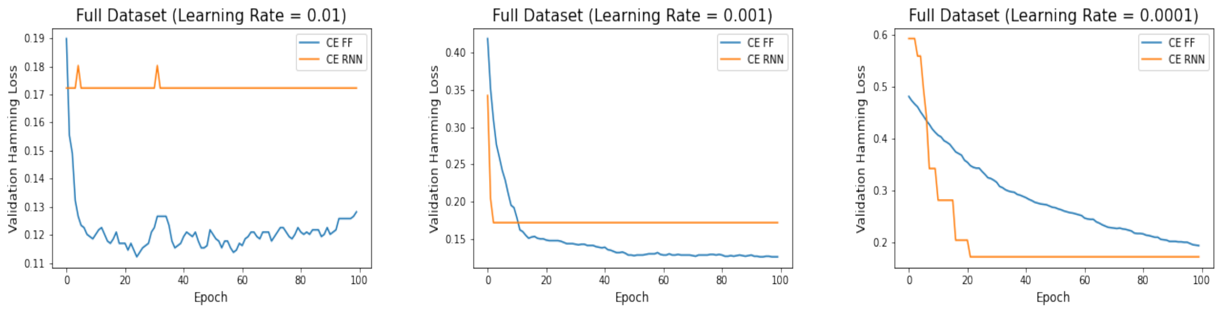


Figure 1: 2 models trained by full dataset

In contrast with constant constant thresholds, this time CE FF seems to have higher hamming loss than CE RNN for all 3 learning rate. It also seems that the learning rate of 0.001 has the lowest hamming loss.

Learning Rate	CE FF	CE RNN
0.01	0.3020833333	0.2379807692
0.001	0.17868589740	0.1730769231
0.0001	0.2115384615	0.1770833333

Table 1: Hamming Loss with Threshold Function Learning

All the 4 models are trained with the reduced dataset using constant threshold and the results are shown in figure 2. Similar to the full dataset case, RNN in general has higher hamming loss than Feed-Forward for learning rate of 0.01 and 0.001. The same conclusion seems to hold for learning rate of 0.0001 if training epoches increase.

From the right and middle sub figures, for both RNN and Feed-Forward network, the hamming loss is quite close between BPMLL and cross entropy. In fact, the left sub figure shows BPMLL has higher hamming loss than cross entropy for both RNN and Feed-Forward network.

Also, smaller learning rate needs more training epoches for finale convergence.

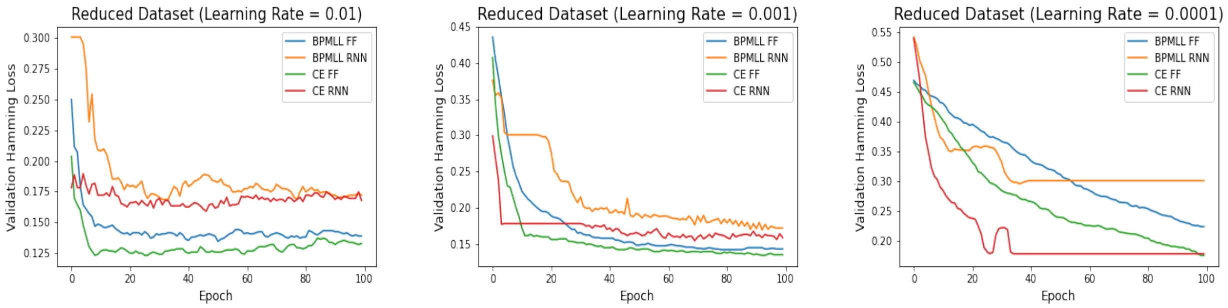


Figure 2: 4 models trained by reduced dataset

Similarly, as shown in Table 2, the 4 models are also trained with Threshold Function Learning, but the hamming loss is not reduced for all 4 models. Also, no clear winner among 4 models is observed.

Learning Rate	CE FF	BPMLL FF	CE RNN	BPMLL RNN

4 Discussion & Conclusions

References

[1] Matthew R. Boutell et al. *Learning multi-label scene classification.* 2004.

[2] André Elisseeff and Jason Weston. "A Kernel Method for Multi-Labelled Classification". In: *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic.* NIPS'01. Vancouver, British Columbia, Canada: MIT Press, 2001, pp. 681–687.

[3] Simon Haykin. *Neural Networks and Learning Machines.* Pearson, 2009.

[4] Mark A. Kramer. "Nonlinear Principal Component Analysis Using Autoassociative Neural Networks". In: *AIChE Journal* 37.2 (1991), pp. 233–243. DOI: https://doi.org/10.1002/aic.690370209.

[5] Oscar Luaces et al. "Binary relevance efficacy for multilabel classification". In: *Progress in Artificial Intelligence* 1.4 (2012), pp. 303–313. DOI: 10.1007/s13748-012-0030-x.

[6] Mohamad Aljnidi Kadan Aljoumaa Maha Alkhayrat. "A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA". In: *Journal of Big Data* 7.9 (2020). ISSN: 2196-1115. DOI: https://doi.org/10.1186/s40537-020-0286-0.

[7] Andrew Kachites McCallum. "Multi-label text classification with a mixture model trained by EM". In: *AAAI 99 Workshop on Text Learning.* 1999.

[8] Xu-Ying Liu Min-Ling Zhang Yu-Kun Li and Xin Geng. "Binary relevance for multi-label learning: an overview". In: *Frontiers of Computer Science* 12.2 (2018), pp. 191–202. DOI: 10.1007/s11704-017-7031-7.

[9] Lan Huong Nguyen and Susan Holmes. "Ten quick tips for effective dimensionality reduction". In: *PLoS computational biology* 15.6 (2019), e1006907.

[10] Gualtiero Piccinini. "The First Computational Theory of Mind and Brain: A Close Look at McCulloch and Pitts' *Logical Calculus of Ideas Immanent in Nervous Activity*". In: *Synthese* 141.2 (2004), pp. 175–215. DOI: 10.1023/B:SYNT.0000043018.52445.3e.

[11] Min-Ling Zhang and Zhi-Hua Zhou. "ML-KNN: A lazy learning approach to multi-label learning". In: *Pattern Recognition* 40.7 (2007), pp. 2038–2048. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2006.12.019. URL: https://www.sciencedirect.com/science/article/pii/S0031320307000027.

[12] Min-Ling Zhang and Zhi-Hua Zhou. "Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization". In: *IEEE Transactions on Knowledge and Data Engineering* 18.10 (2006), pp. 1338–1351. DOI: 10.1109/TKDE.2006.162.