Hybrid Learning Approaches

Semantic Data Web Technologies Seminar Report

Debanjali Biswas, Shyamnath Premnadh, Barshana Banerjee and Mst. Mahfuja Akter

Informatik, Universität Bonn, Germany
{debanjali.biswas, s6shprem, s6babane, s6msakte}@uni-bonn.de

Abstract

This report describes different Hybrid Learning approaches to solve real-world problem scenarios. We consider four problem domains here: (1) Pattern Recognition, (2) Text Extraction, (3) Stock Prediction and (4) Disease Prediction (Medical Domain) and discuss about the performances of the Hybrid Learning algorithms proposed for these domains. This study was a part of the Semantic Data Web Technologies Seminar course held in the Summer Semester 2019 at the University of Bonn.

1 Introduction

Hybrid learning algorithms are a combination of learning paradigms that provide advantages that the original paradigms do not possess. Use of hybrid learning generally has a certain number of benefits over the individual algorithms, which include [5]:

- Better advantages over the individual learning algorithms
- Should be able to solve complex problem that an individual learning algorithm can not solve or is very difficult for it to solve
- Overcome the limitations of each single calibrated algorithm
- The combined algorithms should share the complementary advantages

The report is organized as follows: The next section (Section 2) describes the four problem scenarios considered. Section 3 illustrates the different proposed approaches for the four problem domains mentioned in the previous section including their need for hybridization, their architecture and algorithm. Section 4 discusses the experimental results and evaluations for the four proposed models. The last section (Section 5) presents the discussions and conclusions of the four models.

2 Problem Scenarios

The four problem domains that we consider here are as follows:

2.1 Pattern Recognition

Pattern recognition is the task of classifying data into different classes or categories by using computer algorithms which automatically detect regularities in data.[1] The proposed hybrid model that we consider for this problem is a self-supervised learning algorithm which combines an unsupervised clustering K-Means Fast Learning Artificial Neural Network (KFLANN) with a typical supervised Back-propagation learning algorithm (BP Network).[5]

2.2 Text Extraction

Information Extraction(IE) systems facilitates extracting from the document all the parts that correctly fill a set of predefined output slots. Traditionally, for IE, various machine learning classifiers have been used which divide the documents into fragments, based on the various descriptive features of the document(length, presence of terms etc). The ML Algorithm later associates each of these fragments with a predefined output slot. However these systems classify each fragment independently of the other fragments, thereby missing some important information about the document structure. This is exactly what the proposed hybrid learning approach tries to solve, for semi-structured texts. Information Extraction in a semantic web scenario can help in the extraction and linking of concepts, entities and relations. A hybrid learning approach on top of this will result in more accurate triple formation.

2.3 Stock Prediction

Stock market prediction is considered to be a difficult task in money time-series forecasting. The primary factors behind this is the underlying uncertainties in the market movement including economic, political conditions. But, it has been proven that this market price behaviour is not random at all. It has a dynamic non-linear pattern instead. The proposed hybrid model we consider for this problem is a two step neural network architecture constructed by combining Support Vector Machine (SVM) and Empirical Mode Decomposition (EMD) which is appropriate to predict behaviour of a non linear time series.

2.4 Disease Prediction

Diabetes occurs due to abnormal level of glucose and insulin in human bodies. The major issue with diabetes is the mildness of it's symptoms which delay the diagnosis. The significant progress of disease leads to many other medical complication. The paper under consideration here predicts the onset of diabetes by combining one supervised network paradigm Support vector machine (SVM) and an unsupervised Deep learning network.[3] It also measures the progression of the disease.

3 Methodology

3.1 Pattern Recognition

The proposed hybrid model is a combination of an unsupervised K-means Fast Learning Artificial Neural Network (KFLANN) and a supervised Back-propagation (BP) Network.

3.1.1 Need for Hybridization

Although the KFLANN and BP have a lot of benefits, each possesses a set of drawbacks. In the case of KFLANN, it is an unsupervised learning algorithm. Hence, it is unable to perform interpolation of possible outputs. Thus, the hybridized model optimizes the learning characteristics of the BP and the fast learning and clustering abilities of the KFLANN to perform interpolation activities.

On the other hand, despite the high accuracy of a supervised BP learning algorithm, it is expensive to provide complete target value for most of the real-world problems. To minimize the dependency on the external teacher, KFLANN can provide the platform to reproduce these exemplars and thus creating a network which is trained by another network. Therefore developing a self-supervised learning algorithm.

3.1.2 Hybrid Model Behavior

The goal of the proposed hybrid model is to remove the necessity of manual interrogation of the network by handling the learning of patterns in a self-supervised manner. As KFLANN provides consistent and statistically sound clusters, they are used as actual targets in the training the BP Network. The flowchart of this hybrid model is shown in Figure 1.

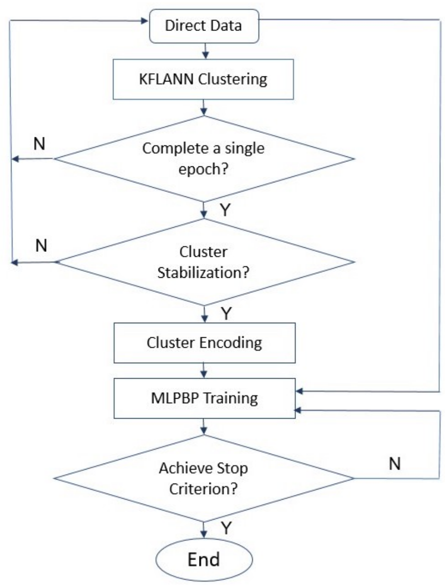

Figure 1: Flowchart of Hybrid KFLANN and MLPBP Model

For cluster membership assignment, the input data is fed forward to the KFLANN. In this process, each data point is allocated to a single cluster. A Cluster Assignment List (CAL) is created after training the KFLANN. MLPBP is trained with the same training data. However, the teacher value is obtained from the CAL. So, the learning process starts with the KFLANN clustering, which prepares the target values for the training data used in MLBP training to obtain generalization. The gradient descent learning process by the BP Network will start after the convergence of the KFLANN centroids and continue until the stopping criterion is attained. The Cluster encoding is done by directly mapping the cluster to the cluster index. For instance, if the cluster index is assigned as k for a particular data point, then it means that particular data pattern belongs to cluster k.

3.2 Text Extraction

The proposed method is a hybrid learning model for Information Extraction on semi-structured documents that combines typical ML based text classification techniques(kNN, Bayes etc) and Hidden Markov Model(HMM). The traditional ML classification algorithm produces a locally optimal classification as they ignore the relationship between the different fragments. By using

HMM(whose input is the result from the ML classifier), we produce a optimal classification considering the relation between the fragments.

3.2.1 Proposed Approach

The proposed approach is divided into two phases, each phase for the two models. The approach is elucidated relative to bibliographic references. This being a semi-structured document the information in the document has a certain ordering to it (although not rigidly enforced), which helps in the extraction process.

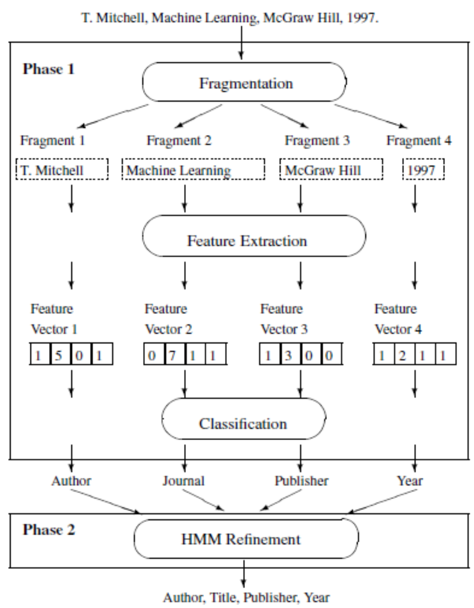

Figure 2: Proposed approach for Hybrid Information Extraction

1. Phase 1-**Extraction using an ML based classifier**
   - Fragmentation: Division of the semi-structured document into fragments, to fill each of the output slots. In the case of bibliographic references these output slots could be the author, title, affiliation, editor, publisher etc. For the purpose of this paper, 14 different slots have been used.
   - Feature Extraction: A feature vector is created for each fragment containing the possible candidates to fill the output slots. Three different feature sets were used to test the proposed model
     (a) Manual1 (20 features defined in [2])
     (b) Manual2 (9 features defined in [4])
     (c) Automatic (a simple set of 100 words chosen from the training corpus based on information gain [6])
   - Classification: In this phase the ML classifier algorithm is applied and does the preliminary assignment of candidates to each output slot. The ML algorithm is trained initially using a corpus of known fragments.

2. Phase 2-**Refinement of results**. The results from the phase 1 are fed into the phase 2. The HMM improves the initial classification produced by the ML algorithm, producing a globally optimal classification.

3.3 Stock Prediction

3.3.1 Proposed Approach

The proposed hybrid model is a combination of Support Vector Machine (SVM) and Empirical Mode Decomposition (EMD).

3.3.2 Need for Hybridization

As the time series data is completely dynamic, using a single SVM cannot give us desired result. This is because it is not capable to capture the inherent structure and relations between input-output data as the number of factors influencing them are always much higher.

3.3.3 Proposed Model Description

The proposed model is divided into 3 stages.The workflow of this hybrid model is shown in Figure 3.

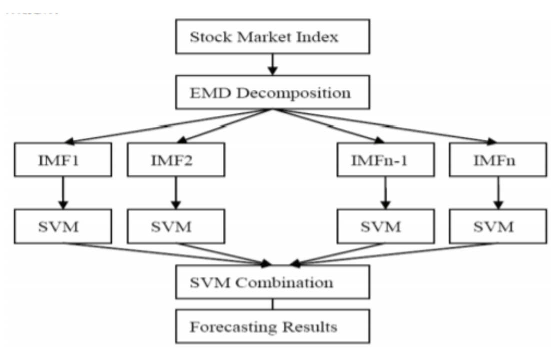

Figure 3: Workflow of EMD-SVM model.

First stage : In the first stage, we take the original dynamic non stationary stock market data and apply empirical model decomposition on it in order to get different intrinsic mode functions which are stationary. So they contain more accurate information about the signal after decomposition.

Second stage : Once we have stationary intrinsic functions, we apply SVM to the model to eliminate unwanted noise impact. We do that by choosing appropriate kernel function and other learning parameters for each intrinsic function. The reason behind using SVM is that it uses a regularization term derived from Structural Risk Minimization Principle.

Final stage : In this stage, we combine all the SVM results from last step in order to get final prediction to forecast the new market index.

3.4 Disease Prediction

The proposed hybrid deep learning method classifies the disease more accurately and compares the efficiency with individual neural network SVM and Deep learning network.

3.4.1 Why Hybridization

All though Support Vector machine can separates data into two clusters by a hyperplane, the classification accuracy is not good, whereas the standalone Deep learning networks gives better performance on large datasets. The purpose of combining these two model is to get higher accuracy than the individual models.

3.4.2 Data process of Hybrid model

The dataset used for disease prediction is known as the Pima Indians dataset which records the female population over 15 years old. There are nine different attributes in this dataset as shown in Table 1 which are used for diabetic prediction and progression. The attribute "Plasma glucose concentration" is used to predict the progression of disease, as set by American Diabetes Association.

Sl no.	Name of Attributes
1	No. of Pregnancies
2	Plasma Glucose Concentration
3	Diastolic Blood Pressure (mm Hg)
4	Triceps Skin Fold Thickness (mm)
5	2-hour Serum Insulin (mu U/ml)
6	Body Mass Index
7	Diabetes Pedigree Function
8	Age
9	Class Variable (0 or 1)

Table 1: Dataset for Hybrid SVM and Deep neural Model

Firstly, the missing data in the dataset is imputed by mean value and then all attributes are normalized to get similar ranged data. After pre-processing of dataset, Support vector machine is applied for classification which is followed by Deep neural learning algorithm. The first layer of Deep learning is replaced by a linear SVM layer. The workflow of this hybrid model is shown in Figure 4. To avoid over fitting, the dataset is divided into training set, cross validation set and test set by 6:2:2.

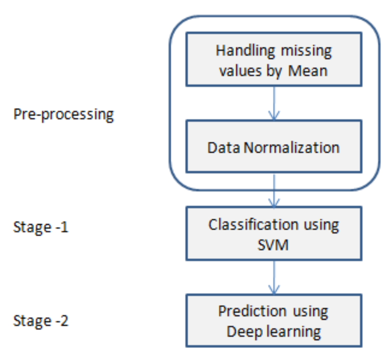

Figure 4: Framework of Hybrid model for disease prediction

4 Evaluation

4.1 Pattern Recognition

4.1.1 Hybrid Network Setting

Standard deviation is assumed as the feature tolerance setting for the KFLANN models in this hybrid approach. For all datasets considered, the Vigilance value is set to 0.9.

The MLP structure is a Multiple Input Multiple Output (MIMO) where, the number of input nodes is equal to the number of features, the number of output nodes is equal to the number of clusters created by the KFLANN and the number of hidden nodes is given by the following heuristic formula:

$$Hidden\,node = (input\,node + number\,of\,classes)/2$$

The stopping criteria is whichever of the following is attained first:

- a maximum number of epochs reached, which is set as 8000
- minimum error obtained, which is set to 0.000001.

4.1.2 Datasets Description

The benchmark datasets,downloadable from UCI Machine Learning Repository, were selected for performance verification of the hybrid model. The dataset description is shown in Table 2. The first column is the name of the dataset and the second column is the number of features. The next two columns gives us the number of instances in the training and test sets respectively. The last column is the number of classes present.

Dataset	Feature	Training	Test	Class
Iris	4	120	30	3
Thyroid	5	140	75	2
Wine	13	114	28	3
WBC	9	614	154	2
Dermatology	34	286	72	6
Zoo	16	82	19	7

Table 2: Dataset Description for Hybrid KFLANN and MLPBP Model

4.1.3 Classification Results

The Receiver Operating Characteristics (ROC) was considered to be the measure to organize the performance. Area Under Curve (AUC), computed by equation (1), was used to measure the prediction accuracy of the hybrid method as the approach is not sensitive to the skewness of class distribution.

$$\begin{aligned} E &= \frac{1}{K}\sum_{m=1}^{K} e_m \\ e_m &= \frac{number\,of\,pattern\,misclassified\,in\,class\,m}{n_m} \\ AUC &= 1 - E \end{aligned} \tag{1}$$

where K is the number of classes, e_m is the classification error of class m and n_m is the number of data patterns in class m.

Table 3 depicts the classification accuracy in percentage of the Hybrid model along with their individual models, MLPBP and KFLANN, on the test data for all the different datasets. As seen in the table, the accuracy of the proposed model is better than the unsupervised KFLANN clustering but not better than the supervised BP learning process. However, the hybrid model has a major advantage over BP that it does not need the class information. It is expensive and difficult to obtain the class labels for some problem domains. Hence, the hybrid model is a cost effective way of solving pattern recognition problems. Experimental results show that in almost all the datasets, the accuracy of the hybrid model is comparable to the BP.

Dataset	**Hybrid**	**MLPBP**	**KFLANN**
Iris	96.97	100.00	90.48
Thyroid	89.93	94.67	78.13
Wine	96.67	92.86	90.77
WBC	93.95	94.03	91.77
Dermatology	83.33	80.88	94.44
Zoo	100.00	100.00	100.00
Average	**93.48**	**93.74**	**90.93**

Table 3: Classification Accuracy for Hybrid KFLANN and MLPBP Model

4.2 Text Extraction

The model was evaluated using a collection of 6000 bibliographic references, with each fragment being mapped to the correct tags (to indicate the output slots). These were further divided into two equal parts of 3000 references each, one of which was used for training and the other for testing the performance of the model. Each feature set (mentioned in Methodology) is evaluated for testing the performance of the model, with and without HMM, for 3 different ML classifiers (kNN, Bayes and Part). The evaluation measure used is **Precision**. Figure 5 shows the Average Precision per reference(with and without HMM) for the different combinations of the feature set and the aforementioned ML Classifiers. It can be inferred from the results of the table that there is a steady increase in the performance of the model with HMM as compared to the model without HMM - this is depicted by the **Gain**.

Feature Set	Classifier	Precision without HMM	Precision with HMM	Gain
Manual1	PART	72.17%	76.40%	4.22%
Manual1	Bayes	66.70%	74.72%	8.01%
Manual1	kNN	71.96%	76.28%	4.32%
Manual2	PART	73.48%	77.29%	3.80%
Manual2	Bayes	69.03%	77.27%	8.23%
Manual2	kNN	76.17%	81.16%	4.99%
Automatic	PART	49.91%	72.45%	22.54%
Automatic	Bayes	50.11%	68.25%	18.14%
Automatic	kNN	51.47%	73.57%	22.10%

Figure 5: Results obtained for hybrid Information Extraction on the test corpus

4.3 Stock Prediction

4.3.1 Datasets Description

Shanghai Composite Index from Wind database is used as dataset for this experiment. The dataset covers from the period of November 27, 2008 to February 22, 2010.

4.3.2 Performance evaluation

From Figure 6 we can clearly say that Shanghai Composite Index is not stationary and it can affect performance of the prediction. Figure 7 shows partitioned dataset after applying EMD where these resulting intrinsic mode functions becomes stationary. Hence, results more accurate prediction.

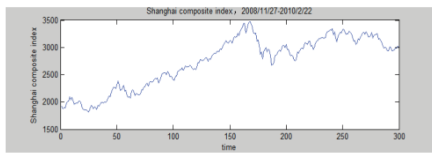

Figure 6: The trend of Shanghai Composite Index.

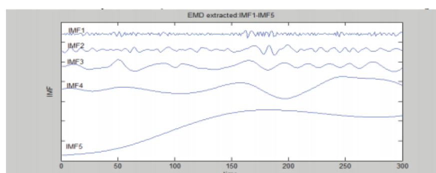

Figure 7: IMFs of Shanghai Composite Index through EMD.

	SVM Prediction Results			EMD-SVM Prediction Results		
Actual Data	Prediction	Absolute Error	Comparative Error	Prediction	Absolute Error	Comparative Error
2982.58	3030.20	47.62	1.60%	2957.15	25.43	0.85%
3022.18	3025.52	3.34	0.11%	2974.68	47.50	1.57%
3060.62	3004.67	55.95	1.83%	3010.18	50.44	1.65%
3051.94	2968.92	83.02	2.72%	3048.86	3.08	0.10%
3087.84	2922.53	165.31	5.35%	3076.79	11.05	0.36%
3073.11	2871.43	201.68	6.56%	3089.97	16.86	0.55%
3097.01	2821.76	275.25	8.89%	3090.45	6.57	0.21%
3023.37	2778.49	244.87	8.10%	3081.10	57.73	1.91%
3031.07	2744.68	286.39	9.45%	3064.34	33.27	1.10%
3053.23	2721.29	331.94	10.87%	3042.51	10.72	0.35%

Figure 8: Comparative predictions between SVM and hybrid EMS-SVM model.

To measure the actual performance enhancement of this hybrid EMD-SVM model as compared to single SVM model, absolute error and comparative error have been used as evaluation metrics. Figure 9 gives the comparative prediction results between SVM and EMD-SVM models. It is clear that hybrid model outperforms single SVM prediction by minimizing the error from the range 0.11

4.4 Disease Prediction

4.4.1 Confusion Matrix

Reprocessed (imputed and Normalized) data is fed into SVM, Deep Neural Network and Hybrid Deep learning method. Table 5 shows the correctly and incorrectly predicted data for above mentioned method. Using those data, performance and efficiency is measured and compared to each other.

	True positive	**False positive**	**True negative**	**False negative**
SVM	140	46	454	128
Deep Neural Network	150	50	450	118
Hybrid Deep learning Network	187	70	430	81

Table 4: Confusion matrix of Deep neural, SVM and Hybrid deep neural

4.4.2 Performance evaluation

Performance Table 5 shows the comparison of performances on the dataset, which is calculated from correctly classified and misclassified data for SVM , Deep learning and Hybrid Deep learning algorithm. It compares the Accuracy, Area Under Curve (AUC), Mean Square Error (MSE) and Precision. Hybrid learning approach has better accuracy and precision value and also the least MSE compared to the other individual methods. The area under ROC curve is shown Figure 9 also indicates the efficiency of method, higher the AUC value better the performance.

	Deep Neural Network	**SVM**	**Hybrid Deep learning Network**
Accuracy	0.7812	0.7734	0.8034
AUC	0.8538	0.7152	0.8733
MSE	0.1489	0.2266	0.1376
Precision	0.7502	0.7526	0.7625

Table 5: Performance analysis of Deep neural, SVM and Hybrid deep neural

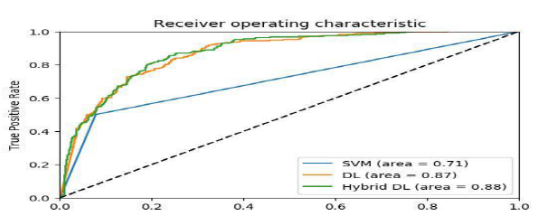

Figure 9: Comparing ROCs of Deep neural, SVM and Hybrid deep neural

The comparison chart (Table 5) and the area under Receiver Operating characteristic (ROC) curve (Figure 9) are clearly indicating that the hybrid method provides most satisfactory performance in every evolutionary aspects.

5 Conclusion

The hybrid model for **Pattern Recognition** is providing an automated means to train the supervised BP network using the output of the KFLANN as the teacher values. Experimental results support that the performance of the Hybrid method is comparable with the individual algorithms. The benefits of the Hybrid over the individual models, MLPBP and KFLANN, are as follows:

- The hybrid model is a cost effective learning model without any requirement of external class information
- It provides a platform for the KFLANN to share interpolation capabilities with BP
- The performance of classification is not compromised with the self-supervised training

The hybrid model used for **Information Extraction on Semi-structured documents** clearly delineates that the use of HMM offsets the comparatively low performance of using just the ML classifier on the feature sets. A better precision is obtained on all the feature sets when the hybrid model is used. The novelty of this model is that it combines two techniques that was not yet combined for IE. The proposed model can be extended to other domains by changing the definition of the HMM structures (in our model they hidden states were considered to be connected to each other), and by using different machine learning strategies to generate the input fragments.

The hybrid model for **Stock prediction** is providing an automated way to predict future stock. Though the proposed algorithm or the paper under consideration shows a great improvement, there is still a lot of scope for further work in this area. In this social media age, public emotion can make a market upside down just with a single easy tweet. So, it is time to introduce other non-technical details (Eg. Sentiment analysis) as input to the in neural network. We can say that, merging all these inherent factors in this model will enhance the performance to a certain extent without any doubt.

The hybrid model used in **Disease Prediction**, where the first layer of Deep learning network is replaced by linear SVM layer has comparably better performance and is more efficient than any of individual methods. The current dataset is only focusing on female population. It is not a good idea for a method to work only on specific gender data, it is intended to check this method for male population also. Although the proposed method measures the disease prediction and progression, still there is some possibility to get misclassified data. The performance of decision support system can be optimized by training with more extensive data.

6 Acknowledgement

I am indebted to my supervisor Dr. Maria Maleshkova for guiding us throughout the duration of the Semantic Data Web Technologies Seminar course. I sincerely thank her for giving us this opportunity and acknowledge her valuable support and reviews for the successful completion of this course.

References

[1] Christopher M Bishop. *Pattern recognition and machine learning*. Information science and statistics. Springer, New York, NY, 2006. Softcover published in 2016.

[2] Carla C. R. Nunes and Flvia Barros. Prodext: A knowledge-based wrapper for extraction of technical and scientific production in web pages. pages 106–115, 01 2000.

[3] Deeksha Kaul, Harika Raju, and B.K. Tripathy. Comparative analysis of pure and hybrid machine learning algorithms for risk prediction of diabetes mellitus. 07 2017.

[4] Remco R. Bouckaert. Low level information extraction: a bayesian network based approach. 12 2002.

[5] Wong Lai Ping and A. T. L. Phuan. A novel hybrid learning scheme for pattern recognition. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2099–2104 vol. 4, July 2005.

[6] Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 412–420, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.