A Survey on Machine Learning Approach in Data Mining

M.Deepthi[1] T. Poongothai[2]
[1]M.Tech Student [2]Associate Professor
[1,2]Department of Computer Engineering
[1,2]K.S.R College of Engineering, Department of Information Technology, Tiruchengode, TamilNadu, India

Abstract— Machine learning is a branch of computer science which evolves from the pattern recognition and computational learning theory. The task of machine learning is categorized into supervised learning, unsupervised learning and reinforcement learning. In this paper, techniques such as support vector machine to extract the substring and for named entities, OSC-NMF for pattern, PCA to detect cancer disease, decision tree to diagnosis heart disease, margin-based censored regression approach, novel multi-task learning techniques are surveyed.

Key words: Support Vector Machine, OSC-NMF, PCA, Decision Tree, Margin-based Censored Regression Approach, Novel Multi-Task learning techniques

I. INTRODUCTION

In company data set was increasing day by day. Data mining is the branch of computer science used to extract important data from large data set. Machine learning is one for the technique used in data mining for pattern recognition and computational learning theory. Machine learning has three categorized – supervised learning, unsupervised learning, reinforcement learning.

- Supervised learning: It is the machine learning task of inferring a function from labeled training data.
- Unsupervised learning: In this labels are not given to the learning algorithm; instead find its own structure.
- Reinforcement learning: A program interacts with a dynamic environment in its goal, without a teacher explicitly telling it whether it achieve its goal or not. Each learning has different learning approach and each one is applied for different concept in this paper.

II. LITERATURE SURVEY

A. Extracting key-Substring-Group Feature for Text Classification (2006)

As compared to previous research study on generative Markov chain models, discriminative machine learning methods like support vector machine (SVM) is successful in text classification with word features. It is neither effective nor efficient to apply them straightly by taking all substrings in the corpus as features. In this paper, we proposed to partition all substring into statistical equivalence groups, and then pick those groups which are important as features for text classification. In this method, suffix tree based algorithm that can extract such features in linear time. By experimenting, SVM with key-substring-group features can achieve best performance for various text classification tasks.

B. Towards Heterogeneous Temporal Clinical Event Pattern Discovery: A Convolutional Approach (2012)

A convolutional approach which uses nonnegative matrix factorization based framework for open-ended temporal pattern discovery over large collections of clinical records is proposed in this paper and calls this method as One- Sided Convolutional NMF (OSC-NMF). It presents how to adapt OSC-NMF to extract patterns from one data samples. The experimental results on data sets are presented to demonstrate the effectiveness of the proposed approaches.

C. Recognizing Medication related entities in Hospital Discharge Summaries using Support Vector Machine (2010)

The i2b2 NLP challenge is a task to extract six types of medication related named entities (NEs), such as medication names, mode, dosage, duration and reason from hospital discharge summaries. At the first step, a machine learning algorithm such as conditional random field (CRF) and maximum entropy have been applied to the named entity recognition (NER) task. In this paper, we presented a support vector machine (SVM) based method to recognize medication related entities and it achieved the best F-score of 90.05%.

D. Using Deep Learning to Enhance Cancer Diagnosis and Classification (2013)

Unsupervised feature learning is used for cancer detection and type analysis from gene expression data. The advantage of this method is to apply data from various types of cancer to automatically form features which help to enhance the detection and diagnosis of a specific one. In the proposed method, this uses PCA to address the very high dimensionality of the initial raw feature space followed by sparse feature learning techniques. This method not only shows that it can be used to improve the accuracy in cancer classification problems, but also demonstrates that it provides a more general and scalable approach to deal with gene expression data across various cancer types.

E. Using Decision Tree for Diagnosing Heart Disease Patients (2011)

In this paper, decision tree is one of the data mining techniques used and most research has applied J4.8 decision tree based on Gain Ratio and binary discretization, a widely used benchmark data set is used. The two other successful types of decision trees are Gini Index and Information Gain that are less used in the diagnosis of heart disease. The sensitivity, specificity, and accuracy are calculated to evaluate the performance of the alternative decision tree and the highest accuracy is 79.1% by the equal width discretization information gain decision tree. The research presented a model that performs J4.8 decision tree and bagging algorithm in the diagnosis of heart disease patients. From the results it is concluded that most of the researchers in the diagnosis of heart disease by using binary discretization with gain ratio decision tree, applying multi-interval equal frequency discretization with 9 voting gain ratio decision trees provides outperforming results in the diagnosis of heart disease patients.

F. An Integrated Machine Learning Approach to Stroke Prediction (2010)

In this paper, for stroke prediction the Cox proportional hazard model with a machine learning approach on the CHS dataset are compared. This feature selection algorithm may not work well in other dataset with correlated features and it evaluates the performance of each feature individually. To overcome this problem, an L1 regularized feature selection algorithm to crop the features to apply for conservative mean feature selection for fine-tuning before it is used. In this paper, an integrated machine learning approach is presented to combine the elements of data imputation, prediction and feature selection and comparing with the cox proportional hazards model and that the machine learning methods will outperform the cox model in terms of binary stroke prediction and stroke risk estimation. By calculating the conservative mean for feature selection which gives us the better performance as compared to other methods. We present a novel prediction algorithm, margin-based censored regression that achieves a best concordance index than the cox model. Further, without performing clinical trials, our method can be used for identifying potential risk factors for diseases. It will motivate the application of machine learning methods in healthcare field.

G. Modeling Disease Progression via Fused Sparse Group Lasso (2012)

In this paper, Novel multi-task learning techniques are developed to find the disease progression measured by cognitive scores and select biomarkers prediction of the progression. In multi-task learning, cognitive scores is considered as a task, and multiple prediction tasks at various time points are performed to capture the temporal smoothness of the models across different time points. A novel convex fused sparse group Lasso formulation that allows the simulation selection of a common set of biomarkers for various time points and specific sets of biomarkers for various time points using the sparse group Lasso penalty and include the temporal smoothness using the fused Lasso penalty. It is challenging to solve by using several non-smooth penalties. The main technique of this paper is it exhibits a decomposition property along with the proximal operator associated with the proposed formulation and can be computed efficiently. To further improve the model, two non-convex formulations are proposed, which are expected to reduce the shrinkage bias in the convex formulation and it will employ the difference of convex (DC) programming technique. Results show that the proposed progression models are more effective than an existing multi-task learning formulation for disease progression.

III. Conclusion

By surveying the different machine learning approach, it is concluded that the techniques used in each paper gives the better performance result. Support Vector Machine gives the best outcome for text classification and medical related entities. The experimental results on both synthetic and real world data sets are presented by OSC-NMF to demonstrate the effectiveness of the system. Decision tree provides best results in the diagnosis of heart disease patients. As a conclusion, machine learning approach is improving its performance in health care environment.

References

[1] D. Zhang and W. S. Lee, “Extracting key-substring-group features for text classification,” in Proc. 12th ACM SIGKDD Conf. Knowl. Discovery Data Mining, 2006, pp.474-483.

[2] F. Wang, N. lee, J.Hu, J. sun and S. Ebadollahi, “Towards heterogeneous temporal clinical event pattern discovery: A convolutional approach,” in proc. 18th ACM SIGKDD Conf. Knowl. Discovery Data Mining, 2012, pp.453-461.

[3] S. Doan and H. Xu, “Recognizing medication related entities in hospital discharge summaries using support vector machine,” in Proc. Int. Conf. Comput. Linguistics, 2010, pp.259-266.

[4] R. Fakoor, F. Ladhak, A. Nazi and M. Huber, “Using deep learning to enhance cancer diagnosis and classification,” presented at the Int. Conf. mach. Learn., Atlanta, GA,USA,2013.

[5] M. Shouman, T. Turner and R, Stocker, Using decision tree for diagnosing heart disease patients,” in Proc. 9th Australasian Data Mining Conf., 2011, pp.23-30.

[6] A. Khosla, Y. Cao, C. C-y. lin, H-K Chiu, J.Hu, and H. Lee, “ An integrated machine learning approach to stroke prediction,” in proc. 16th ACM SIGKDD Conf. Knowl. Discovery Data mining, 2010, pp. 183-192.

[7] J. Zhou, J. Liu, V. A. Narayan and J. Ye, “Modeling disease progression via fused sparse group lasso,” in Proc. ACM SIGKDD Conf. Knowl. Discovery Data Mining, 2012, pp. 1095-1103.