ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 10, October 2018

A Survey on CDPCF: Concise Discriminative Patterns Based Classification Framework

Ashwini Shahpurkar[1], Prof. S B Chaudhari[2]

Department of Computer Engineering, Savitribai Phule University of Pune,

JSPM's Jayawantrao Sawant College of Engineering, Hadapsar, Pune[1,2]

Abstract: Example based arrangement was initially proposed to enhance the precision utilizing chosen visit designs, where numerous endeavours were paid to prune a colossal number of non-discriminative successive examples. Then again, tree-based models have demonstrated solid capacities on numerous arrangement errands since they can undoubtedly fabricate high-arrange communications between various highlights and furthermore handle both numerical and downright highlights and high dimensional highlights. Simple models such as generalized linear models have ordinary performance but strong interpretability on a set of simple features. There are different series which includes tree-based models, organize numerical, categorical and high dimensional features into a comprehensive structure with rich interpretable information in the data. Frequent pattern-based classification methods have shown to be very effective at classifying categorical or high dimensional sparse datasets

Keywords: Discriminative pattern, Tree based model

I. INTRODUCTION

Various algorithms and models have been introduced for classification. Generalized linear classification models, such as support vector machine and logistic regression, usually have reasonably good performance but lack the power of modeling complex high-order interactions between features. Tree-based models, such as random forest [1] and gradient boosted trees have been deployed in many practical settings and often achieved high accuracy, because the high model complexity of trees provides the chance of high-order combinations of different features. Neural network is another kind of powerful classifiers, especially in image classification problems, which models nonlinear relationship among features and usually performs with high prediction accuracy. However, in real world applications, many people favour generalized linear models instead of complex models, including trees and neural networks, as long as the accuracies are enough in practice, because they are mature, flexible, more efficient when making prediction, and easier to be understood by providing probabilistic interpretation. The low interpretability makes complex models not suitable for many applications, such as classification problems in medical applications and scientific domains, in which feature importance and contribution of feature combinations from the model could be highly useful for obtaining intuitive understanding of the application. The ultimate goal would be to construct accurate models that are also simple enough to interpret.

To address this challenge, one possible solution is to feed constructed high-order features to generalized linear models and enhance accuracy and interpretability. Along this direction, many previous pattern-based models have been established in the last decade and demonstrated powers in several domains, including (1) association rule-based classification on categorical data [4]; (2) frequent pattern-based classification on text and graph [6] data; (3) discriminative pattern-based classification on general data [2, 3], which mine discriminative patterns starting with frequent patterns and have shown their advantages over both tree-based models and generalized linear models. Many efforts are paid to prune a huge number of non-discriminative frequent patterns in those models; however, the number of extracted patterns utilized in later classification models is at the magnitude of thousands, which is still large.

This synopsis proposes a novel discriminative patterns-based classification framework (CDPCF) with the goal to generate a very concise high-order classification model. The key component of CDPCF is a fast and effective pattern extraction algorithm. Instead of starting with frequent patterns, it first train tree-based models to generate a large set of hypothetical high-order patterns, and then it investigate all prefix ways from root hubs to leaf hubs in the tree-based models as our discriminative examples. Finally, it further compresses the number of discriminative patterns by selecting the most effective pattern combinations that fit into a generalized linear model with high classification accuracy.

In this way, CDPCF generates a set of discriminative high order patterns with high productivity and interpretability. From another perspective, it can view CDPCF as a way to compress the multi-tree based models by only selecting the most discriminative pattern combinations and fitting them into a generalized linear model. Surprisingly, CDPCF

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 10, October 2018

achieves comparable or even improved performance over the original tree-based models with only storing dozens of robust discriminative patterns. Such models can also be extremely useful for applications (e.g., mobile apps), where model storage and online computational cost are restricted.

- To tackle existing disadvantages, this system proposed a novel concise discriminative patterns-based classification framework (CDPCF) with the goal to generate a very concise high-order classification model.
- The key component of CDPCF is a fast and effective pattern extraction algorithm. Instead of starting with frequent patterns, it first train tree-based models to generate a large set of hypothetical high-order patterns, and then it explore all prefix paths from root nodes to leaf nodes in the tree-based models as our discriminative patterns.
- Finally, it further compresses the number of discriminative patterns by selecting the most effective pattern combinations that fit into a generalized linear model with high classification accuracy.

II. RELATED WORK

In this paper [1], they propose a systematic framework for frequent pattern-based classification and give theoretical answers to several critical questions raised by this framework. The significant improvement is achieved in classification accuracy using the frequent pattern-based classification framework. In light of this methodology, combined with a proposed include determination calculation, discriminative continuous examples can be produced for building top notch classifiers. A compelling and productive component determination calculation is proposed to choose an arrangement of continuous and discriminative examples for characterization. It achieves good scalability and high accuracy in classifying large datasets. Visit designs are valuable by mapping the information to a higher dimensional space. Each time, one part is used for test and the other parts are used for training. It could harm the classification accuracy due to over fitting. Frequent pattern-based classification methods [2] have shown to be very effective at classifying categorical or high dimensional sparse datasets. They proposed a direct discriminative pattern mining approach DDPMine which directly mines the discriminative patterns and integrates feature selection into the mining framework. They introduce a "feature-centered" mining approach that generates discriminative patterns sequentially on a progressively shrinking FP-tree by incrementally eliminating training instances. It tackles the efficiency issue easily arising from the two-step approach. Associative classification methods in terms of both accuracy and efficiency.
Problem size reduces is slow process. It makes system complex.

Their two step approach [3], which combines random forest and a stepwise selection, provides a realistic approach for selecting an optimal set of features within a reasonable computational time. They also assessed the prediction performance dependency on the initial state of the stepwise selection. Finally, domain linker predictions that consider only local sequence characteristics cannot yield a perfect predictor. DROP (Domain linker pRediction using OPtimal features) performances were superior to previously developed domain linker predictors trained without systematic optimization of the features. They used fast two-step feature selection approach. A useful approach for improving predictions of properties. DROP of features not typically used for boundary prediction. Features that decreased the prediction performances. A significant problem [4] in object detection and classification is that non-linear methods are too expensive for many tasks, while the classes are not well separated by linear hyperplanes in existing feature spaces. They approach this problem introducing an intermediate mapping step where examples are mapped from a given feature space to one where they are easier to separate using a linear classifier. They propose a sparse tree-based mapping method that learns a mapping of the feature vector to a space where a linear hyperplane can better separate negative and positive examples. Combining features from several mappings increases performance. Results mean location when utilized together with a complex straight model. Conceptual dissimilarities result in quite different decision boundaries. Rulefit is a method that does not explicitly create a feature mapping.

They present a comprehensive empirical study [5] of algorithms for fitting generalized additive models (GAMs) with spline and tree based shape functions. Also introduce a new GAM method based on gradient boosting of size-limited bagged trees that yields significantly more accuracy than previous algorithms on both regression and classification problems. The bias-variance analysis that explains how different shapes models influence the additive model. They apply these methods to six classification and six regression tasks. Best method on low- to medium-dimensional datasets. The objective of this work is to build exact models. Characterization issues while holding the comprehensibility of GAM models. They can be easily interpreted by users. They present a framework [6] called GA2M (Generalized Additive 2 Models) for building intelligible models with pairwise interactions. Adding pairwise connections to customary GAMs holds comprehensibility, while considerably expanding model exactness. They propose a novel method called FAST that efficiently measures the strength of all potential pairwise interactions. Therefore we propose a framework using greedy forward stage wise selection strategy to build the most accurate model in H. The build models those are more powerful than GAMs. For simplicity and without loss of generality use this approach. Accuracy drops as the number of feature pairs grows. Fake cooperation's might be accounted for over low thickness locales.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 10, October 2018

The two best calculations outflanked a technique [7] outlined by the test coordinators and also forecasts by ALS clinicians. The DREAM-Phil Bowen ALS Prediction Prize4Life challenge additionally recognized a few potential nonstandard indicators of sickness movement including uric corrosive, creatinine and shockingly, circulatory strain, revealing insight into ALS pathobiology. The test brought about the accommodation of 37 novel calculations from which two winning passages were recognized. It provides both the most robust and the most reliable predictions. It had potential to reduce the population size needed to measure a drug effect by 20%. It has more costly clinical trials. Clinician showed substantially less correlation to the true rate. They presented [8] a global refinement algorithm to improve the fitting power of a pre-trained random forest. They also developed a global pruning algorithm to reduce the over-fitting risk as well as the model size. To address the issues, they propose two techniques, global refinement and global pruning, to improve a pre-trained random forest. The enhanced irregular timberland accomplishes better exactness and littler model size, contrasted with standard arbitrary backwoods and some best in class variations. Optimization is efficient and the testing speed is as fast. The refined model has better execution and littler stockpiling cost. The approach is not suitable for random forest. It does not effectively minimize the global training loss. They focus on the validity [9] of predictive models for monitoring health status, including both model performance and model interpretation. Also, the tree-based models: decision tree and DPClass; provide significant insights into how demographics interact in model prediction. They not only provide promising solutions to monitor health status by simply carrying a smartphone, but also demonstrate how demographics influence predictive models of cardiopulmonary disease. Higher predicting power and better interpretable mechanism. This gives us a potential solution to automatically detect different demographic cohorts. They won't be able to obtain a daily status and risk change for all patients that much efficiently. Population measurement of health status from carried phones with less accuracy. Their results show that [10] a large, deep CNN is capable of achieving record-breaking results on a highly challenging dataset using purely supervised learning. It is prominent that their system's execution corrupts if a solitary convolutional layer is expelled. To lessen overfitting in the completely associated layers they utilized an as of late created regularization strategy called "dropout" that turned out to be extremely successful. Their network takes between 5 and 6 days to train on two GTX 580 3GB GPUs. It reduces overfitting in the fully connected layers. Very efficient GPU implementation of the convolution operation. They did not pre-process the images in any other way. It seems, by all accounts, to be excessively costly for enormous neural systems.

III. PREDICTION FRAMEWORK

The prediction of system [11] reliability and availability requires that the R&A requirements are derived in a specified way. In this paper, a framework was defined for comparing existing reliability and availability analysis methods from the software architecture point of view. Unwavering quality and accessibility must be designed into programming from the beginning of its advancement, and potential issues must be distinguished in the beginning times. The goal is to find which techniques are reasonable for the dependability and accessibility expectation of the present complex frameworks. To break down the unwavering quality of segment based applications as an element of their segments and interfaces.
In this paper [12], they surveyed a large number of existing computational algorithms for miRNA target predictions. The survey is carried out according to the two categories of the target prediction algorithms - the rule-based and the data driven approaches. Particularly, they provided a mathematical definition and formulated the problem of target prediction under the framework of statistical classification.

Finally, they tested a few different algorithms on a set of experimentally validated true miRNA-target pairs and a set of false miRNA-target pairs. In this paper [13], they have presented a novel benchmark framework for software defect prediction. The framework involves evaluation and prediction. In the assessment arrange, diverse learning plans are assessed and the best one is chosen. At that point, in the forecast organize, the best learning plan is utilized to construct an indicator with every single chronicled datum and the indicator is at last used to anticipate imperfection on the new information. The deformity indicator manufactures models as per the assessed learning plan and predicts programming deserts with new information as per the built model. In order to demonstrate the performance of the proposed framework used both simulation and publicly available software defect data sets.

IV. CONCISE DISCRIMINATIVE PATTERN

In this paper [14], they have developed a new framework for mining association rules based on the minimal predictive rules (MPR) concept. They showed that their method can produce a small set of predictive rules. In particular, each govern in the outcome is essential since it briefly depicts an unmistakable example that can't be clarified by some other run in the set. Their experiments on several synthetic and UCI datasets demonstrate the advantage of our framework by returning smaller and more concise rule sets than the other existing association rule mining methods. They call these rules the class association rules. To achieve the goal, first introduce the concept of the minimal predictive rules (MPR).

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 10, October 2018

This review paper [15] addresses all the major aspects of an object detection framework. These include feature selection, learning model, object representation, matching features and object templates, and the boosting schemes. This examination gives a starter, succinct, yet entire foundation of the protest recognition issue. The researchers can choose a framework suitable for their own specific object detection problems and further optimize the chosen framework for better accuracy in object detection. The techniques that utilization edge-based component compose extricate the edge guide of the picture and recognize the highlights of the question as far as edges. The main challenge in the development of the algorithm [16] for gaining high performance to enhance graph mining process is graph isomorphism which is the most costly step since it is an NP-complete problem. Because of expanding size and computational multifaceted nature of example in PC sciences the requirement for productive diagram mining calculation is expanding. This paper researches on correlation of diagram digging calculations and methods for finding the continuous examples. Visit design mining (FPM) is a critical piece of diagram mining that finds designs that reasonably speak to relations among discrete elements.

V. DISCRIMINATIVE PATTERN- BASED PREDICTION FRAMEWORK (DPPRED)

Frequent pattern-based classification methods [17] have shown to be very effective at classifying categorical or high dimensional sparse datasets. In this study, they proposed a direct discriminative pattern mining approach DDPMine which directly mines the discriminative patterns and integrates feature selection into the mining framework. They examine the efficiency issue that arises from the two-step mining framework and propose a direct mining solution. DDPMine accomplishes requests of size speedup with no minimization of arrangement exactness.

In this paper [18], they propose an effective and concise discriminative pattern-based classification framework (DPClass) to address the general classification problem and provide interpretability by incorporating a limited number of discriminative patterns. Comprehensive experiments have demonstrated that DPClass is able to model high-order interactions and present a small amount of interpretable patterns to help human experts understanding. DPClass could increase far and away superior precision by just utilizing best 20 discriminative examples. Moreover, they further compress the number of discriminative patterns by selecting the most effective pattern combinations that fit into a generalized linear model. In this thesis [19], an effective and concise discriminative pattern-based learning framework (DPLearn) is proposed to address the general classification and regression problems and provide interpretability by incorporating a limited number of discriminative patterns. Additionally, the quantity of discriminative examples is additionally compacted by choosing the best example mixes that fit into a summed up straight model. DPLearn first trains tree-based models to generate a large set of hypothetical high-order patterns, and then all prefix paths from root nodes to leaf nodes in the tree-based models are extracted as discriminative patterns.

CONCLUSION AND FUTURE WORK

This synopsis proposed an effective and Concise Discriminative Pattern-Based Classification Framework (CDPCF) to address the general classification problem and provide interpretability by incorporating a limited number of discriminative patterns. CDPCF first extracts the prefix paths from root nodes to non-leaf nodes in tree-based models as candidate discriminative patterns and then further compress the number of discriminative patterns by selecting the most effective pattern combinations according to their predictive accuracy in a generalized linear model. Comprehensive experiments have demonstrated that CDPCF is able to model high-order interactions and present a small amount of interpretable patterns to help human experts understanding the classification tasks. Moreover, it provides comparable or even better accuracy than the previous state-of-the-art pattern-based classification model and the uncompressed random forest model. In future, we plan to extend our CDPCF to a uniform machine learning framework CDPLearn, which supports multi-classes classification, regression, and ranking along the same discriminative pattern selection direction. Another possible direction is to apply CDPCF to labeled textual and sequential data targeting on finding interesting patterns.

ACKNOWLEDGEMENT

I profoundly grateful to **Dr. S B Chaudhari** for his/her expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion. I would like to express my deepest appreciation towards Principal, **Dr. M G Jadhav** HOD department of computer engineering, **Dr. S B Chaudhari** and PG coordinator, **Prof. M D Ingale.** I must express my sincere heartfelt gratitude to all staff members of computer engineering department who helped me directly or indirectly during this course of work. Finally, I would like to thank my family and friends, for their precious support.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 10, October 2018

REFERENCES

[1]. H. Cheng, X. Yan, J. Han, and C.-W. Hsu. "Discriminative frequent pattern analysis for effective classification". In Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on, pages 716– 725. IEEE, 2007.
[2]. H. Cheng, X. Yan, J. Han, and P. S. Yu. "Direct discriminative pattern mining for effective classification". In Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on, pages 169–178. IEEE, 2008.
[3]. T. Ebina, H. Toh, and Y. Kuroda. "Drop: an svm domain linker predictor trained with optimal features selected by random forest". Bioinformatics, 27(4):487–494, 2010.
[4]. M. Kobetski and J. Sullivan. "Discriminative tree-based feature mapping". Intelligence, 34(3), 2011.
[5]. Y. Lou, R. Caruana, and J. Gehrke. "Intelligible models for classification and regression". In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012.
[6]. Y. Lou, R. Caruana, J. Gehrke, and G. Hooker. "Accurate intelligible models with pairwise interactions". In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 623–631. ACM, 2013.
[7]. R. K¨ uffner, N. Zach, R. Norel, J. Hawe, D. Schoenfeld, L. Wang, G. Li, L. Fang, L. Mackey, O. Hardiman, et al. "Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression". Nature biotechnology, 33(1):51–57, 2015.
[8]. S. Ren, X. Cao, Y. Wei, and J. Sun. "Global refinement of random forest". In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 723–730, 2015.
[9]. Q. Cheng, J. Shang, J. Juen, J. Han, and B. Schatz. "Mining discriminative patterns to predict health status for cardiopulmonary patients". In Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB '16, pages 41–49, New York, NY, USA, 2016. ACM.
[10]. A. Krizhevsky, I. Sutskever, and G. E. Hinton. "Imagenet classification with deep convolutional neural networks". In Advances in neural information processing systems, pages 1097–1105, 2017.
[11]. Anne Immonen and Eila Niemelä, "Survey of reliability and availability prediction methods from the viewpoint of software architecture," Regular Paper on 12 January 2007.
[12]. Dong Yue1, Hui Liu2 and Yufei Huang, "Survey of Computational Algorithms for MicroRNA Target Prediction," Current Genomics, Vol. 10, No. 7 479, in 2009.
[13]. Qinbao Song, Zihan Jia, Martin Shepperd, Shi Ying, and Jin Liu, "A General Software Defect-Proneness Prediction Framework," IEEE Transactions On Software Engineering, Vol. 37, No. 3, May/June 2011.
[14]. Iyad Batal and Milos Hauskrecht, "A Concise Representation of Association Rules using Minimal Predictive Rules," IEEE Conference paper, volume 6321, 2010.
[15]. Dilip K. Prasad, "Survey of The Problem of Object Detection In Real Images," International Journal of Image Processing (IJIP), Volume (6), Issue (6), 2012.
[16]. Harsh J. Patel, Rakesh Prajapati, Prof. Mahesh Panchal, Dr. Monal J. Patel, "A Survey of Graph Pattern MiningAlgorithm and Techniques," IJAIEM Volume 2, Issue 1, January 2013.
[17]. Hong Cheng, Xifeng Yan, Jiawei Han, Philip S. Yu, "Direct Discriminative Pattern Mining for Effective Classification," IEEE 24th International Conference on Data Engineering, 25 April 2008.
[18]. Jingbo Shang, Wenzhu Tong, Jian Peng and Jiawei Han, "DPClass: An Effective but Concise Discriminative Patterns – Based Classification Framework," International Conference on Data Mining, June 2016.
[19]. Wenzhu Tong, "Dplearn: An Effective but Concise Learning Framework based On Discriminative Patterns," 31 Oct 2016.