Project Title : Big Data Mining and Classification of Intelligent Material

Science Data Using Machine Learning

Author Name : Swetha Chittam,Balakrishna Gokaraju

Year of Publish : 2021

Abstract :

There is a high need for a big data repository for material compositions and their derived analytics of metal strength, in the material science community. Currently, many researchers maintain their own excel sheets, prepared manually by their team by tabulating the experimental data collected from scientific journals, and analyzing the data by performing manual calculations using formulas to determine the strength of the material. In this study, we propose a big data storage for material science data and its processing parameters information to address the laborious process of data tabulation from scientific articles, data mining techniques to retrieve the information from databases to perform big data analytics, and a machine learning prediction model to determine material strength insights. Three models are proposed based on Logistic regression, Support vector Machine SVM and Random Forest Algorithms. These models are trained and tested using a 10-fold cross validation approach.The Random Forest classification model performed better on the independent dataset, with 87% accuracy in comparison to Logistic regression and SVM with 72% and 78%, respectively

Project Title : Underwater target recognition methods based on the framework of deep learning: A survey

Author Name : Bowen Teng and Hongjian Zhao

Year of Publish : 2020

Abstract :

The accuracy of underwater target recognition by autonomous underwater vehicle (AUV) is a powerful guarantee for underwater detection, rescue, and security. Recently, deep learning has made significant improvements in digital image processing for target recognition and classification, which makes the underwater target recognition study becoming a hot research field. This article systematically describes the application of deep learning in underwater image analysis in the past few years and briefly expounds the basic principles of various underwater target recognition methods. Meanwhile, the applicable conditions, pros and cons of various methods are pointed out. The technical problems of AUV underwater dangerous target recognition methods are analyzed, and corresponding solutions are given. At the same time, we prospect the future development trend of AUV underwater target recognition.

Project Title : Predicting drug properties with parameter-free machine learning: pareto-optimal embedded modeling (POEM)

Author Name : Andrew E Brereton,Stephen MacKinnon

Year of Publish : 2020

Abstract :

The prediction of absorption, distribution, metabolism, excretion, and toxicity (ADMET) of small molecules from their molecular structure is a central problem in medicinal chemistry with great practical importance in drug discovery. Creating predictive models conventionally requires substantial trial-and-error for the selection of molecular representations, machine learning (ML) algorithms, and hyperparameter tuning. A generally applicable method that performs well on all datasets without tuning would be of great value but is currently lacking. Here, we describe pareto-optimal embedded modeling (POEM), a similarity-based method for predicting molecular properties. POEM is a non-parametric, supervised ML algorithm developed to generate reliable predictive models without need for optimization. POEM’s predictive strength is obtained by combining multiple different representations of molecular structures in a context-specific manner, while maintaining low dimensionality. We benchmark POEM relative to industry-standard ML algorithms and published results across 17 classifications tasks. POEM performs well in all cases and reduces the risk of overfitting.

Project Title : House Price Prediction Using Machine Learning Algorithm

Author Name : Akshata Ashok Kudavkar, Manali Namdev Nhavkar

Year of Publish : 2019

Abstract :

The real estate market is one of the most competitive in terms of prices and it tends tovary significantly based on a lot of factors,hence it becomes one of the prime fields to apply the concepts of machine learning.Therefore in this project, we present various algorithms while predicting house prices with good accuracy. We tested a regression models such as Simple Linear Regression, Ridge Regression, Lasso Regression, Support Vector Regression, Random Forest Regression, Decision Tree Algorithm and selected the best fit among the algorithm. This project directs us that it can be best application of machine learning models in order to optimize the result .

Project Title : Characterization and identification of lysine glutarylation basedon intrinsic interdependence between positions in the substrate sites

Author Name : Kai-Yao Huang1,2, Hui-Ju Kao

Year of Publish : 2019

Abstract **:**

Background: Glutarylation, the addition of a glutaryl group (five carbons) to a lysine residue of a protein molecule,is an important post-translational modification and plays a regulatory role in a variety of physiological and biological processes. As the number of experimentally identified glutarylated peptides increases, it becomes imperative to investigate substrate motifs to enhance the study of protein glutarylation. We carried out a bioinformatics investigation of glutarylation sites based on amino acid composition using a public database containing information on 430 non-homologous glutarylation sites.Results: The TwoSampleLogo analysis indicates that positively charged and polar amino acids surrounding glutarylated sites may be associated with the specificity in substrate site of protein glutarylation. Additionally, the chi-squared test was utilized to explore the intrinsic interdependence between two positions around glutarylation sites. Further, maximal dependence decomposition (MDD), which consists of partitioning a large-scale dataset into subgroups with statistically significant amino acid conservation, was used to capture motif signatures of glutarylation sites. We considered single features, such as amino acid composition (AAC), amino acid pair composition (AAPC), and composition of k-spaced amino acid pairs (CKSAAP), as well as the effectiveness of incorporating MDD-identified substrate motifs into an integrated prediction model. Evaluation by five-fold cross-validation showed that AAC was most effective in discriminating between glutarylation and non-glutarylation sites, according to support vector machine (SVM). Conclusions: The SVM model integrating MDD-identified substrate motifs performed well, with a sensitivity of 0.677, aspecificity of 0.619, an accuracy of 0.638, and a Matthews Correlation Coefficient (MCC) value of 0.28. Using an independent testing dataset (46 glutarylated and 92 non-glutarylated sites) obtained from the literature,we demonstrated that the integrated SVM model could improve the predictive performance effectively,yielding a balanced sensitivity and specificity of 0.652 and 0.739, respectively. This integrated SVM model has been implemented as a web-based system

Project Title : Competitive Deep-Belief Networks for Underwater

Acoustic Target Recognition

Author Name : Honghui Yang , Sheng Shen

Year of Publish : 2018

Abstract **:**

Underwater acoustic target recognition based on ship-radiated noise belongs to the small-sample-size recognition problems. A competitive deep-belief network is proposed to learn features with more discriminative information from labeled and unlabeled samples. The proposed model consists of four stages: (1) A standard restricted Boltzmann machine is pretrained using a large number of unlabeled data to initialize its parameters; (2) the hidden units are grouped according to categories, which provides an initial clustering model for competitive learning; (3) competitive training and back-propagation algorithms are used to update the parameters to accomplish the task of clustering; (4) by applying layer-wise training and supervised fine-tuning, a deep neural network is built to obtain features. Experimental results show that the proposed method can achieve classification accuracy of 90.89%, which is 8.95% higher than the accuracy obtained by the compared methods. In addition, the highest accuracy of our method is obtained with fewer features than other methods

Project Title : Ensembles of Regularized Linear Models

Author Name : Anthony Christidis,Laks V.S. Lakshmanan

Year of Publish : 2019

Abstract :

We propose an approach for building ensembles of regularized linear models byoptimizing an objective function that encourages sparsity within each model and di-versity among them. Our procedure works on top of a given penalized linear regression estimator (e.g., Lasso, Elastic Net, SCAD) by \_tting it to possibly overlapping sub- sets of features, while at the same time encouraging diversity among the subsets, to reduce the correlation between the predictions from each \_tted model. The predic- tions from the models are then aggregated. For the case of an Elastic Net penalty and orthogonal predictors, we give a closed form solution for the regression coe\_cients in each of the ensembled models. We prove the consistency of our method in possibly high-dimensional linear models, where the number of predictors can increase with the sample size. An extensive simulation study and real-data applications show that the proposed method systematically improves the prediction accuracy of the base linear estimators being ensembled. Possible extensions to GLMs and other models are discussed.

ProjectTitle **:** Predicting drug side effects by multi-labellearning and ensemble learning

AuthorName **:** Wen Zhang, Feng Liu

Year of Publish : 2021

Abstract **:**

Background: Predicting drug side effects is an important topic inthedrug discovery. Although several machine learning methods have been proposed to predict side effects, there is still space for improvements. Firstly, the side effect prediction is a multi-label learning task, and we can adopt the multi-label learning techniques for it. Secondly,drug-related features are associated with side effects, and feature dimensions have specific biological meanings.Recognizing critical dimensions and reducing irrelevant dimensions may help to reveal the causes of side effects.Methods: In this paper, we propose a novel method ‘feature selection-based multi-label k-nearest neighbor method’ (FS-MLKNN), which can simultaneously determine critical feature dimensions and construct high-accuracy multi-label prediction models.Results: Computational experiments demonstrate that FS-MLKNN leads to good performances as well as explainable results. To achieve better performances, we further develop the ensemble learning model by integrating individual feature-based FS-MLKNN models. When compared with other state-of-the-art methods, the ensemble method produces better performances on benchmark datasets. Conclusions: In conclusion, FS-MLKNN and the ensemble method are promising tools for the side effect prediction. The source code and datasets are available in the Additional file 1.