Optimization techniques for feature selection in classification

[1]Dr. K. James Mathai, [2]Kshiti Agnihotri
[1] Professor, [2]Student,
[1]Department of Computer Technology and application
[1]National Institute of Technical Teacher Training & Research, Bhopal, India

***Abstract*—the feature selection process is considered a problem of global combinatorial optimization technique that aims to reduce the number of features, removes irrelevant, noisy and redundant data and results in standard classification accuracy. Feature selection plays key role in machine learning, pattern classification and data mining applications. Therefore, a good feature selection method is needed based on the number of features investigated for sample classification in order to speed up the processing, to reduce the time complexity ,to rate predictive accuracy, and to reduce computational complexity. Although a large body of research has delved into this problem, there is a paucity of survey that indicates trends and directions. This paper attempts to categorize the prevalent popular optimization techniques in feature selection that deals with enhancing classification performance in terms of accuracy and efficiency.**

Key words — Classification, Feature selection, Dimensionality reduction, Feature subset.

I. INTRODUCTION

Data mining techniques are used to handle the massive amount of data that is generally produced from online and offline sources. Hence the storage of information has also become inconvenient due to large datasets[r] that are occupied by high dimensions and high instances. We have 2d possible solutions for the given problem having d features which makes exhaustive search impractical for high dimension feature space. Feature selection process is used to find the most attributive subset of features that leads to agreeable recognition rates for classifiers[r]. The task of feature selection is considered as optimization problem[r] in which the fitness function[r] can be the measure of feature space separately.

Data classification

Classification is a data mining technique[r] used to predict group membership for data instances. It comes under the data analysis process that deals with the class labels of different data forms and help to discover valid patterns and relationships in large data set. Such analysis can help us to provide better understanding of the large data and results into better prediction models. The classification predicts categorical (discrete, unordered) labels and capable of processing a wider variety of data than regression. There are many traditional classification methods like decision tree induction, k-nearest neighbor classifier, Bayesian networks, support vector machines, rule based classification, case-based reasoning, fuzzy logic techniques, genetic algorithm, rough set approach and so on.

Feature Selection

Feature selection is the main technique to reduce the dimensionality reduction [,r,r,r] that has a common goal which is to reduce the number of harmful, redundant and noisy features in a dataset for smooth and fast data processing purposes. The common method of reducing the dimensionality of the data to be analyzed is to reduce the number of features or variables to a more manageable number.

Feature selection or variable selection consists of reducing the available features to a set that is optimal or sub-optimal and capable of producing results which are equal or better to that of the original set. Reducing the feature set scales down the dimensionality of the data which in turn reduces the training time of the induction algorithm selected and computational cost, improves the accuracy of the classification model and makes the data mining results easier to understand and more applicable. The performance of most classification algorithms can be increased by reducing the feature set especially for K-NN algorithm, it may also lower the accuracy of decision trees. Decision trees have the capability of reducing the original feature set in the tree building process, beginning the process with fewer features may affect final performance. Dash and Liu (1997) broke down the feature selection process into four steps: generation, evaluation, stopping criterion, and validation

Feature selection for classification

Supervised learning is used in majority of classification problems where the underlying class probabilities are unknown and each instance is associated with a class label. An irrelevant feature affect the learning process and a redundant feature does not add anything new to the target concept[r].In many classification problems, it is difficult to learn good classifiers before removing these unwanted features due to the huge size of the data. The running time of the learning algorithms can be effectively reduced by reducing the number of irrelevant/redundant features which yield to a more general classifier. This helps in getting a better insight into the underlying concept of a real-world classification problem. Feature selection mainly affects the training phase of classification. After generating features, instead of processing data with the whole features to the learning algorithm directly, feature selection for classification will first perform feature selection to select a subset of features and then process the data with the selected features to

the learning algorithm. The feature selection phase might be independent of the learning algorithm, like filter models, or it may iteratively utilize the performance of the learning algorithms to evaluate the quality of the selected features, like wrapper models. With the finally selected features, a classifier is induced for the prediction phase.

Usually feature selection for classification attempts to select the minimally sized subset of features according to the following criteria,

• The classification accuracy does not significantly decrease

• The resulting class distribution, given only the values for the selected features, is as close as possible to the original class distribution, given all features.

II. RELATED WORK

In this section we describe Algorithms for flat features and Heuristic approaches for feature selection techniques proposed by different authors are discussed.

Algorithm for flat features

Filter Method- The filter based method is based on ranking techniques. The variables that are assigned with a score having an appropriate ranking criterion and the variables which are having score below threshold value are removed. The filter-based approaches offer more generality as they are independent of the supervised learning algorithm. Filter methods are computationally cheaper, avoids over fitting but these methods ignore dependencies between the features. Hence, the selected subset might not be optimal and a redundant subset might be obtained. The best example is Relief that was developed with the distance-based metric function that weights each feature based on their relevancy (correlation) with the target-class. However, Relief is ineffective as it can handle only the two-class problems and also does not deal with redundant features. The modified version of the Relief known as Relief can handle the multi-class problems and deal with incomplete and noisy datasets too. However, it fails to remove the redundant features. Holte developed a rule-based attribute selection known as One R which forms one rule for each feature and selects the rule with the smallest error. Yang & Moody proposed a joint mutual information-based approach (JMI) for classification. It calculates the joint mutual information between the individual feature and the target-class to identify the relevant features, and a heuristic search is adopted for optimization when the number of features is more. The features containing similar information and lesser relevancy to the target-class are treated as redundant features that are to be eliminated. Examples include the Chi squared test, information gain and correlation coefficient scores. [4][5]

Wrapper method - The wrapper method seems to be a "brute force" method. Wrapper methods use heuristics of the learning algorithm and the training set that are better in defining optimal features rather than simply relevant features. Wrapper methods consider the selection of a set of features as a search problem, where different combinations are prepared, evaluated and compared to other combinations. The method uses backward elimination process to remove the insignificant features from the subset. Needs some predefined learning algorithm to identify the relevant feature. The over fitting of feature is avoided using the cross validation. Though wrapper methods are computationally expensive and take more time compared to the filter method, they give more accurate results than filter model. In filter model, optimal features can be obtained rather than simply relevant features. Another advantage is it maintains dependencies between features and feature subsets**. [4][5]**

Embedded method- Embedded method is also called hybrid model as it is a combination of filter and wrapper method. The method attempts feature selection by using the part of learning process of the supervised learning algorithm. The embedded method can be splitted into three categories namely pruning method, built-in mechanism and regularization models. In the pruning method, firstly all the features are processed into the training phase for building the classification model and the features having less correlation coefficient value are removed recursively using the support vector machine (SVM) [r].In the built-in mechanism-based feature selection method, features are selected by using a part of the training phase of the C4.5 and ID3 supervised learning algorithms. In the regularization method, objective functions are used to minimize the fitting error and the features with near zero regression coefficients are eliminated. [4][5]

Heuristic approaches for feature selection

Ideally, feature selection methods search through the subsets of features and try to find the best one among the competing 2m candidate subsets according to some evaluation functions. However this procedure is exhaustive as it tries to find only the best one. It may be too costly and practically prohibitive, even for a medium-sized feature set size. Other methods based on heuristic or random search methods attempt to reduce computational complexity by compromising performance. These methods need a stopping criterion to prevent an exhaustive search of subsets .Hence, here we provide:

A binary cuckoo search and its application for feature selection

[1] Binary Cuckoo Search algorithm has been introduced which is based on the behavior of cuckoo birds, in which the search space is modeled as a d-cube, where d stands for the number of features. In traditional CS, the solutions are updated in the search space towards continuous-valued positions. But in BCS the search space is modeled as a n dimensional Boolean lattice, in which the solutions are updated across the corners of a hypercube for feature selection. It is modeled with different transfer functions that map constant solutions to binary solutions to binary ones. A solution binary vector is employed to select the feature set or not where 1 corresponds whether a feature will be selected to compose the new dataset and 0 otherwise In addition, the Optimum Path Classifier accuracy is used as fitness function. We evaluate the strength of BCS to accomplish the feature selection task comparing it with the binary versions of Bat Algorithm, Firefly Algorithm, Particle Swarm Optimization .The provided simulations and analysis over four public datasets, employ traverse validation strategy to verify how the techniques work for feature selection purposes. The results confirmed that Cuckoo Search has good capabilities to find the best one on two out of four datasets.

Multi-objective genetic algorithm approach to feature subset

[3] The Setback of feature selection is multi-objective in nature and hence optimizing feature subsets with respect to any single evaluation criteria is not sufficient. Thus in order to combine several feature selection criteria the multi-objective optimization of quality subsets apply Multi –Objective Genetic Algorithm. The results confirm that the proposed system is able to determine diverse optimal feature subsets that are well spread in the overall feature space and the classification accuracy of the ensuing feature subsets is reasonably high. The diverse solutions satisfy reasonable levels of optimality with respect to predictive power, non-redundancy and cardinality of feature subsets. These solutions provide the users with several choices of feature subsets

Bat algorithm and feature selection based on rough set

[2] A new selection technique based on rough-sets and bat algorithm is proposed. Bat Algorithm (BA) is a feature selection approach in which bats fly within feature subset space and they discover best feature combinations. Bat algorithm requires only primitive and simple mathematical operator's not composite operators such as crossover and mutation. It is computationally inexpensive in terms both memory and runtime .A fitness incorporates both the classification accuracy and the number of selected features and hence balances the classification performance and reduction size. The used rough-set based fitness function ensures enhanced classification result keeping also minor feature size Rough set theory provides an arithmetical tool to find out data dependencies and reduce the number of features included in dataset by purely structural method. Rough sets have been an advantage tool for medical applications. The complete solution to detect nominal reduce is to produce all possible reduces and choose one with negligible cardinality which can be done by constructing kind of discernibility function from the dataset and simplifying it.

Strengths

- Computationally inexpensive in terms of memory and runtime
- Rough set has been advantageous in reducing the number of features

Binary ant colony optimization for feature selection

Ant colony optimization techniques have been widely used to solve the combinatorial optimization problems. By keeping this in mind, ABACO has been introduced to solve the feature selection problems. A graph model is generated by treating the features as graph nodes and is completely connected to each other. In the graph, each node has two sub nodes that represent selecting and deselecting of features. Ant should visit all the features in ACO algorithm to select the nodes. At the end, each ant carries a binary vector of the same length as the number of features, where 1 implies selecting and 0 implies deselecting of the corresponding feature. The experimental comparison verifies that algorithm provides good classification accuracy using a small feature set than another existing selection ACO –based feature selection method.

Strengths

- Search capability is higher in the setback space
- Good at finding the minimal feature subset.

Cuckoo optimization algorithm for feature selection in high dimensional datasets

[6] Binary Particle Swarm Optimization (BPSO) is used for feature selection. Two Chaotic maps are embedded in binary particle swarm optimization a logistic map and tent map. The purpose of chaotic map is to determine inertia weight of BPSO. A Chaotic Binary Particle Swarm Optimization (CBPSO) is used a method to implement feature selection and k-nearest neighbor method with leave one out traverse validation serves as a classifier to evaluate the classification accuracies. Here a wrapper model for feature selection is adopted. The two chaotic maps show different dynamic behavior. The behavior affects the search capacity of CBPSO. The two different chaotic sequences for the inertia value are applied to the feature selection process. The fallout show that CBPSO with tent map obtained higher accuracy than CBPSO with a logistic map**.**

Feature selection method using genetic algorithm for the classification of small and high dimensional data

[7]An competent feature selection method that finding and selecting informative measures from small or high dimension data which maximum the classification accuracy use genetic algorithm to search out and identify the potential informative features combination for classification and then use the classification accuracy from the support vector machine classifier to decide the fitness in genetic algorithm. Chromosome representation is used in genetic algorithm. A bit value of 1 in the chromosome means the feature is included in specified subset and bit value of 0 means the corresponding feature is not included in the specified subset. The main components of genetic algorithm are feature subset selection and SVM as classifier. The model of the chromosome representation in the proposed approach reduces the combination number of feature subsets with fitting chromosome length. The model further decreases the complexity searching on feature database. The results show the selected features are good to get high classification accuracy for training data of small or high dimension data.

III. DISCUSSION AND CHALLENGES

There are different challenges and concerns that can be mentioned about feature selection for classification

Scalability and Stability

The scalability of current algorithms can be jeopardize with the tremendous growth of dataset sizes especially when used in online classifier. The scalability of feature selection algorithms is a big problem as they require sufficient number of samples to obtain. The storage of full dimensionality of the features in the memory and the requirement of iterative process where each sample is visited

more than once until convergence are the big challenges of the scalability of the feature selection process that are being trying to solve by optimization techniques.

IV.CONCLUSION AND FUTUREWORK

This paper represents the need for feature selection in classification. The majority of the classification requires supervised learning process that helps in prediction model. The unwanted features of huge size of data affect the running time of learning algorithm and by reducing those redundant features yield good classifier. The optimization techniques are used to find out the best among the possible combinations of the feature subsets that helps to reduce the computational complexity. The exhaustive search of feature subsets can be optimized by certain meta-heuristic approaches that helps to find out the global best among search criteria.

REFERENCES

[1] L.A.M.Pereira, D.Rodrigues, T.N.S.Almeida, C.O.Ramos ,A.N.Souza ,X.S.Yang and J.P.Papa, "A Binary Cuckoo search for feature Selection" ,Springer International Publishing Switzerland 2014.

[2] R. Y. M. Nakamura, L. A. M. Pereira, K. A. Costa, D. Rodrigues, J. P. Papa, X.-S. Yang, "BBA : A Binary Bat Algorithm for feature selection", 2012 XXV SIBGRAPI Conference on Graphics, Patterns and Images

[3] Newton Spolaor, Ana Carolina Lorena, Huei Diana Lee "Use of Multi objective Genetic Algorithms in Feature Selection, 2010 Eleventh Brazilian Symposium on Neural Networks

[4] Xing Liu, Lin Shang "A Fast Wrapper Feature Subset Selection Method based On Binary Particle Swarm Optimization" 2013 IEEE Congress on Evolutionary Computation.

[5] D. Asir Antony Gnana Singh, S. Appavu alias Balamurugan, E. Jebamalar Leavline "Literature Review on Feature Selection Methods for High-Dimensional Data", International Journal of Computer Applications (0975 – 8887) February 2016

[6] Suresh Dara, Haider Banka" A Binary PSO Feature Selection Algorithm for Gene Expression Data, International Conference on Advances in Communication and Computing Technologies,2014

[7] Shivani Shrivastri, Rahul Deshmukh "Data Classification Particle Swarm Optimization and Gravitational Search Algorithm" International Journal of Innovative Research in Science, Engineering and Technology", February 2014