International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 2

A Survey on Classification of Feature Selection Strategies

R. Pradeepa
Research Scholar, A.V.C College of Arts and Science
Mannampandal, Mayiladuthurai, Tamil Nadu, India

K. Palanivel
Professor, A.V.C College of Arts and Science
Mannampandal, Mayiladuthurai, Tamil Nadu, India

ABSTRACT

Feature selection is an important part of machine learning. The Feature selection refers to the process of reducing the inputs for processing and analysis, or of finding the most meaningful inputs. A related term, feature engineering (or feature extraction), refers to the process of extracting useful information or features from existing data.

Mining of particular information related to a concept is done on the basis of the feature of the data. The accessing of these features hence for data retrieval can be termed as the feature extraction mechanism. Different type of feature extraction methods is being used. In this paper, the different feature selection methodologies are examined in terms of need and method adopted for feature selection. The three types of method are mainly available, such as Shannon's Entropy, Bayesian with K2 Prior and Bayesian Dirichlet with uniform prior (default). The objectives of this survey paper is to identify the existing contribution made by using their above mentioned algorithms and the result obtained.

Keywords: *Feature selection, Feature selection methods, Bayesian classification, Preprocessing ideas and Decision tree*

1. INTRODUCTION

Feature selection is critical to building a good model for several reasons. One is that feature selection implies some degree of cardinality reduction, to impose a cutoff on the number of attributes that can be considered when building a model. Data almost always contains more information than is needed to build the model, or the wrong kind of information. For example, you might have a dataset with 500 columns that describe the characteristics of customers; however, if the data in some of the columns is very sparse you would gain very little benefit from adding them to the model, and if some of the columns duplicate each other, using both columns could affect the model.

1.1 FEATURE SELECTION

Machine learning works on a simple rule. This becomes even more important when the number of features is very large.

Due to increasing demands for dimensionality reduction, research on feature selection has deeply and widely expanded into many fields, including computational statistics, pattern recognition, machine learning, data mining, and knowledge discovery. Highlighting current research issues, Computational Methods of Feature Selection introduces the basic concepts and principles, state-of-the-art algorithms, and novel applications of this tool. A feature selection can be seen as the combination of a search techniques for proposing new feature subsets. Along with an evaluation measure which score the different feature subsets.

The three main categories of feature selection methods are wrapper method, Filter method, and embedded method.

1.2 IMPORTANCE OF FEATURE SELECTION

Feature selection is always performed before the model is trained. With some algorithms, feature selection techniques are "built-in" so that irrelevant columns are excluded and the best features are automatically discovered. Each algorithm has its own set of default techniques for intelligently applying feature reduction. However, you can also manually set parameters to influence feature selection behavior.

During automatic feature selection, a score is calculated for each attribute, and only the attributes that have the best scores are selected for the model. You can also adjust the threshold for the top scores. Data Mining provides multiple methods for calculating these scores, and the exact method that is applied in any model depends on these factors:

The algorithm used in your model

- The data type of the attribute
- Any parameters that you may have set on your model

Feature selection is applied to inputs, predictable attributes, or to states in a column. When scoring for feature selection is complete, only the attributes and states that the algorithm selects are included in the model-building process and can be used for prediction. If you choose a predictable attribute that does not meet the threshold for feature selection the attribute can still be used for prediction, but the predictions will be based solely on the global statistics that exist in the model.

1.3 FEATURE SELECTION ALGORITHMS

There are three general classes of feature selection algorithms:

- Filter methods,
- Wrapper methods
- Embedded methods.

1.3.1 FILTER METHODS

Filter feature selection methods apply a statistical measure to assign a scoring to each feature. The features are ranked by the score and either selected to be kept or removed from the dataset. The methods are often univariate and consider the feature independently, or with regard to the dependent variable.

Some examples of some filter methods include the Chi squared test, information gain and correlation coefficient scores.

1.3.2 WRAPPER METHODS

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. A predictive model us used to evaluate a combination of features and assign a score based on model accuracy.

The search process may be methodical such as a best-first search, it may stochastic such as a random hill-climbing algorithm, or it may use heuristics, like forward and backward passes to add and remove features.

An example if a wrapper method is the recursive feature elimination algorithm.

1.3.3 EMBEDDED METHODS

Embedded methods learn which features best contribute to the accuracy of the model while the model is being created. The most common type of embedded feature selection methods are regularization methods.

Regularization methods are also called penalization methods that introduce additional constraints into the optimization of a predictive algorithm (such as a regression algorithm) that bias the model toward lower complexity (fewer coefficients).

Examples of regularization algorithms are the LASSO, s Net and Ridge Regression.

2. LITERANER SURVEY

Sadri. S et.al, (2015)., Diabetes is one of the most prevalent diseases in the world today with high mortality and morbidity rate, thus one of the biggest health problems in the world. There are many ways to diagnose the disease; one of these methods is data mining algorithms. The use of data mining on medical data has brought about important, valuable, and effective achievements, which can enhance the medical knowledge to make necessary decisions. The diagnosis of type II diabetes through data mining algorithms. In this article, Naive Bayes, RBF

Network, and J48 are the data mining algorithms used to diagnose type II diabetes. The so-called algorithms perform diagnosis using Weka.

Shivakumar B. L, Alby. S, (2014)., Data mining would be a valuable asset for diabetes researchers because it can unearth hidden knowledge from a huge amount of diabetes-related data. Various data mining techniques help diabetes research and ultimately improve the quality of health care for diabetes patients. This paper provides a survey of data mining methods that have been commonly applied to Diabetes data analysis and prediction of the disease. This paper describes some ideas for a dependable Semantic Web. Semantic Web is a technology for understanding Web pages. It is important that the Semantic Web is secure. In addition, data exchanged by the Web has to be of high quality. The processes that the Web supports have to meet certain timing constraints. This paper discusses these aspects, and describes how they provide a dependable Semantic Web.

Md Rafiul. H, et.al, (2016)., This paper investigates the effectiveness of a hybrid genetic and fuzzy set algorithm for the recognition of gait patterns in falls risk patients. In a previous work, we have shown the usefulness of fuzzy set techniques for gait pattern identification. In this paper, we apply a genetic algorithm in conjunction with fuzzy logic rules to better select the optimal combination of pathological gait features for improved gait diagnostic capability. The genetic algorithm was introduced in order to select the optimum combination of gait features. Using cross validation test data, the results indicated that the generalization performance, in terms of accuracy, for the hybrid system was 97.5%, compared to 9.3% that was obtained using only the fuzzy system. The generalization performance of the gait classifier was also analyzed by determining the areas under the receiver operating characteristic plot.

Ranjit Abraham et.al, (2016)., Naive Bayes classifier has gained wide popularity as a probability-based classification method despite its assumption that attributes are conditionally mutually independent given the class label. This paper makes a study into discretization techniques to improve the classification accuracy of Naive Bayes with respect to medical datasets. Our experimental results suggest that on an average, with minimum description length (MDL) discretization the Naive Bayes classifier seems to be the best performer compared to popular variants of Naive Bayes as well as some popular non-Naive Bayes statistical classifiers.

Cristina Oprea, (2014)., In recent years, there has been an explosion in the rate of using technology that help discovering the diseases. For example, DNA microarrays allow us for the first time to obtain a "global" view of the cell. It has great potential to provide accurate medical diagnosis, to help in finding the right treatment and cure for many diseases. Various classification algorithms can be applied on such micro-array datasets to devise methods that can predict the occurrence of Leukemia disease. In this study, we compared the classification accuracy and response time among eleven decision tree methods and six rule classifier methods using five performance criteria. The experiment results show that the performance of Random Tree is producing better result. Also it takes lowest time to build model in tree classifier. The classification rules algorithms such as nearest- neighbor-like algorithm (NNge) is the best algorithm due to the high accuracy and it takes lowest time to build model in classification.

Sujata. D, Bichitrananda. P, Tripathy. B. K, (2012)., A major challenge in biomedical studies in recent years has been the classification of gene expression profiles into categories, such as cases and controls. This is done by first training a classifier by using a labeled training set containing labeled samples from the two populations, and then using that classifier to predict the labels of new samples. Such predictions have recently been shown to improve the diagnosis and treatment selection practices for several diseases. This procedure is complicated, however, by the high dimensionality of the data. While microarrays can measure the levels of thousands of genes per sample, case-control microarray studies usually involve no more than several dozen samples. Standard classifiers do not work well in these situations where the number of features (gene expression levels measured in these microarrays) far exceeds the number of samples. Selecting only the features that are most relevant for discriminating between the two categories can help construct better classifiers, in terms of both accuracy and efficiency. This paper provides a comparison between dimension reduction technique, namely Partial Least Squares (PLS)method and a hybrid feature selection scheme, and evaluates the relative performance of four different supervised classification procedures such as Radial Basis Function Network (RBFN), Multilayer

Perceptron Network (MLP), Support Vector Machine using Polynomial kernel function (Polynomial- SVM) and Support Vector Machine using RBF kernel function (RBF-SVM) incorporating those methods. Experimental results show that the Partial Least-Squares (PLS) regression method is an appropriate feature selection method and a combined use of different classification and feature selection approaches makes it possible to construct high performance classification models for microarray data.

Shaleena. K. P, et.al, (2015)., Predicting student performances in order to prevent or take precautions against student failures or dropouts is very significant these days. Student failure and dropout is a major problem nowadays. There can be many factors influencing student dropouts. Data mining can be used as an effective method to identify and predict these dropouts. In this paper, a classification method for prediction is been discussed. Decision tree classifiers are used here and methods for solving the class imbalance problem is also discussed.

Parthiban .C, Balakrishnan .M, (2016)., This research paper objective is to provide analysis of classifying the data set using novel naïve classification algorithm. Here this algorithm applied on two dataset which are the first one is training data set and other one is test data set, and this experiment attempts to compare the classification and accuracy of the proposed algorithm with the two datasets. Several constraints used for analytical purpose which are classification accuracy, True Positive Rate, False positive Rate sensitivity, Precision, Recall and f-measure using confusion matrix. Finally results are given in the tabular form to facilitate comparative analysis.

3. FEATURE SELECTION ALGORITHM

The feature selection algorithm removes the irrelevant and redundant features from the original dataset to improve the classification accuracy. The feature selections also reduce the dimensionality of the dataset; increase the learning accuracy, improving result comprehensibility. The feature selection avoid over fitting of data. The feature selection also known as attributes selection which is used for best partitioning the data into individual class.

The feature selection method also includes the selection of subsets, evaluation of subset and evaluation of selected feature.

4. FEATURE SELECTION SCORES

The interestingness score is used to rank and sort attributes in columns that contain non-binary continuous numeric data.

Shannon's entropy and two Bayesian scores are available for columns that contain discrete and discretized data. However, if the model contains any continuous columns, the interestingness score will be used to assess all input columns, to ensure consistency.

5. INTERESTINGNESS SCORE

A feature is interesting if it tells you some useful piece of information. However, interestingness can be measured in many ways. Novelty might be valuable for outlier detection, but the ability to discriminate between closely related items, or discriminating weight, might be more interesting for classification.

The measure of interestingness that is used in Data Mining is entropy-based, meaning that attributes with random distributions have higher entropy and lower information gain; therefore, such attributes are less interesting. The entropy for any particular attribute is compared to the entropy of all other attributes, as follows:

Interestingness(Attribute) = - (m - Entropy(Attribute)) * (m - Entropy(Attribute))

Central entropy, or m, means the entropy of the entire feature set. By subtracting the entropy of the target attribute from the central entropy, you can assess how much information the attribute provides.

This score is used by default whenever the column contains nonbinary continuous numeric data.

6. SHANNON'S ENTROPY

Shannon's entropy measures the uncertainty of a random variable for a particular outcome. For example, the entropy of a coin toss can be represented as a function of the probability of it coming up heads.

Analysis Services uses the following formula to calculate Shannon's entropy:

$$H(X) = -\sum P(xi) \log(P(xi))$$

This scoring method is available for discrete and discretized attributes.

7. BAYESIAN WITH K2 PRIOR

Data Mining provides two feature selection scores that are based on Bayesian networks. A Bayesian network is a directed or acyclic graph of states and transitions between states, meaning that some states are always prior to the current state, some states are posterior, and the graph does not repeat or loop. By definition, Bayesian networks allow the use of prior knowledge. However, the question of which prior states to use in calculating probabilities of later states is important for algorithm design, performance, and accuracy.

The K2 algorithm for learning from a Bayesian network was developed by Cooper and Herskovits and is often used in data mining. It is scalable and can analyze multiple variables, but requires ordering on variables used as input. For more information, see Learning Bayesian Networks by Chickering, Geiger, and Heckerman.

This scoring method is available for discrete and discretized attributes.

8. BAYESIAN DIRICHLET EQUIVALENT WITH UNIFORM PRIOR

The Bayesian Dirichlet Equivalent (BDE) score also uses Bayesian analysis to evaluate a network given a dataset. The BDE scoring method was developed by Heckerman and is based on the BD metric developed by Cooper and Herskovits. The Dirichlet distribution is a multinomial distribution that describes the conditional probability of each variable in the network, and has many properties that are useful for learning.

The Bayesian Dirichlet Equivalent with Uniform Prior (BDEU) method assumes a special case of the Dirichlet distribution, in which a mathematical constant is used to create a fixed or uniform distribution of prior states. The BDE score also assumes likelihood equivalence, which means that the data cannot be expected to discriminate equivalent structures. In other words, if the score for If A Then B is the same as the score for If B Then A, the structures cannot be distinguished based on the data, and causation cannot be inferred.

9. CONCLUSION

Among the existing feature selection algorithms, some algorithms involve only in the selection of relevant features without considering redundancy. Dimensionality increases unnecessarily because of redundant features and it also affects the learning performance. And some algorithms select relevant features without considering the presence of noisy data. Presence of noisy data leads to poor learning performance and increases the computational time.

Our study concludes that there is a need for an effective unified framework for feature selection which should involve in the selection of best feature subset without any redundant and noisy data. It should be applied for all types of data and it should also able to scale up with increasing dimensionality

10. REFERENCE:

1) **CRISTINA. O**, "Performance evaluation of the data mining classification methods", Information society and sustainable development, (2014).

2) **Dunne. K el.al,** "Solution to instability problems with sequential wrapper-based approaches to feature selection", Journal Of Machine Learning Research,2002.

3) **Kashif. J et.al,** "Feature Selection based on Class-Dependent Densities for High Dimensional Binary Data", IEEE Transactions on Knowledge and Data Engineering, Vol 24, No 3, 2012

4) **Kohavi. R,** "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid," ser. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. AAAI Press (1996), pp. 202–207.

5) **Liu. H and Motoda. H,** "Feature Selection for Knowledge Discovery and Data Mining", Kluwer Academic Publishers, 1998.

6) **Rajeswari. B and Aruchamy. R,** "Survey On Data Mining Algorithms to Predict Leukemia Types", Ijrset Volume 2, Issue 5, (2010).

7) **Ranjit. A et.al,** "A comparative analysis of discretization methods for Medical Datamining with Naïve Bayesian classifier", International Conference on Information Technology, IEEE 2016.

8) **Shaleena. K. P et.al,** "Data mining techniques for predicting student performance", International Conference on Engineering and Technology, IEEE 2015.

9) **Sujata .D et.al,** "A Hybrid Data Mining Technique for Improving the Classification Accuracy of Microarray Data Set", I.J. Information Engineering and Electronic Business, (2012),2, 43-50.

10) **Zilin .Z et.al,** "Hybrid Feature Selection Method based on Rough Conditional Mutual Information and Naïve Bayesian Classifier", Hindawi Publishing Corporation, ISRN Applied Mathematics, Vol 2014