INTERNATIONAL JOURNAL FOR INNOVATIVE RESEARCH IN MULTIDISCIPLINARY FIELD ISSN: 2455-0620 Volume - 6, Issue - 4, Apr – 2020
Monthly, Peer-Reviewed, Refereed, Indexed Journal with IC Value: 86.87 Impact Factor: 6.497
Received Date: 13/04/2020 Acceptance Date: 28/04/2020 Publication Date: 30/04/2020

Data Analytics with Efficient Mining of Frequent Patterns on Decision Making

[1]S.Rijutha, [2]M. DineshKumar, [3]R. Haripriya,
[1] Assistant professor, [2, 3] UG Scholar
[1, 2, 3] Department of Information Technology, SNS College of Engineering, Coimbatore
Email - [1] rijuthakumar@gmail.com, [2] kumarmdinesh021@gmail.com, [3] raghavan87k@gmail.com

Abstract: *Feature and variable selection have become the most focused areas of application in research for which datasets with many number of variables are available. Feature selection in neural network can select features which are essential and discard unwanted and indifferent features. Such a method may pick up some useful but dependent features, all of which may not be needed. The proposed scheme, named as Feature Selection Multi-layer Perception (FSMLP), uses controlled redundancy for selecting the features both for classification and function approximation or prediction type problems. Several data sets like synthetic data set are used to demonstrate the effectiveness of the algorithms. In order to control the redundancy a measure of linear dependency is considered.*

Nonlinear measures of reliance, such as correlative information are used since they are straightforward. These methods can report for feasible nonlinear fine interactions between tools, as well as that between attributes and the problem being cracked. They can also manage the level of idleness among the selected attributes.

Presently, the amount of huge data stored in educational database is considered. These database contain the useful information for predicting the students performance. The most useful data mining technique used in educational database is the classification technique. In this work, the classification task is used to predict student's final grade. As there are many proposals that are used for data categorization, the decision tree technique is used here.

Key Words: *feature selection, Multilayer perception, function approximation, synthetic dataset.*

1. INTRODUCTION:

In machine learning and statistics, feature selection/variable selection/attribute selection or sometimes variable subset selection, is the process of selecting a subset of pertinent features (variables, predictors) for constructing a useful model. Attribute selection techniques are used mainly for the following three reasons:

1. As it is easier for researches or users to interpret since models are simplified.
2. Reduced training times.
3. Enhanced generality by simplifying over fitting (formally, variance reduction).

The central hypothesis while using a feature selection technique is that the data may contain number of features that are either redundant or extraneous, thus it can be removed without sustaining much loss of information. Needless (redundant) or extraneous (irrelevant) features are two distinct notions, since one pertinent (relevant) feature may be redundant in the presence of another pertinent feature with which it is strongly correlated.

Feature selection techniques should be differentiated from feature extraction. Feature withdrawal creates new features from functions of the original features, at the same time feature selection proceeds with a subset of the selected features. Feature selection techniques are often used in domains where there is number of features and moderately a small number of samples (or data points). In feature selection technique, the analysis of written texts and DNA microarray data are the most important process carried out. It consists of large number of features (thousands), and few samples (tens/hundreds).

2. FEATURE SELECTION AND SUBSET SELECTION:

A feature selection algorithm is usually the combination of a hunt technique for suggesting new feature subsets, along with an estimation measure which scores the diverse feature subsets. The simplest algorithm is used to check each achievable subset of features discovering the one which minimizes the error rate. This is a thorough search of the space, and is computationally inflexible for all but the least feature sets. The choice of assessment metric profoundly sways the algorithm, and it is these assessment metrics which

discern between the three main categories of element assortment (feature selection) algorithms: wrappers, strain (filter) and embedded techniques.

Wrapper methods use an analytical model to score feature subsets. Each new subset is used to guide/train a model, which is tested on a hold-out set. Counting the number of faults made on that present set (the error rate of the model) shows the score for that subset.

Filter methods use a proxy measure instead of the slip rate to score a feature subset. This computation is chosen to be quick to work out, even as capturing the effectiveness of the attribute (feature) set. Familiar measures comprises of correlative (mutual) information, the point wise correlative information, Pearson product-moment correlation coefficient, inter class space or the scores of important tests for each class/feature integrations. Filters are usually less computationally rigorous than wrappers, but they create a feature set which is not tuned to a particular type of analytical model. This lack of tuning means a feature set from a filter is most common than the set from a wrapper, usually giving lower forecast performance than a wrapper.

Subset selection estimates a subset of features as a group for aptness. Subset selection algorithms can be split up into Wrappers, Filters and Embedded. Wrappers use a hunt algorithm to hunt through the space of achievable features and assess each subset by running a model on the subset. Wrappers can be computationally luxurious and have a problem of over fitting to the technique. Filters are alike as wrappers in the search approach, but instead of assessing a model, a trouble-free filter is assessed. Embedded techniques are implanted in and unambiguous to a model.

3. FEATURE SELECTION SYSTEM ANALYSIS:

Feature selection plays a significant role in pattern acknowledgment and system discovery. It is renowned that, for a given problem, all features that typify a data point are not usually of equal significance; some features may have a derogatory influence on the mission at hand. This may be true regardless of whether the problem is of classification, function approximation, or forecast (prediction). Use of more features adds more degrees of freedom to the system, and hence the learning system gets superior freedom to remember the data, which may result in deprived generalization. When more features are used, it may lead to elevated design and decision-making cost. Attribute selection techniques can be classified in diverse ways. One frequently used classification method clusters the feature selection methods into filter methods and wrapper methods. The filter methods do not require any feedback from the classifier or the predictor, which finally use the selected features. However, a wrapper method assesses the utility of the features using the classifier (or the predictor), which finally uses the selected features.

4. FSMLP AND ID3 ALGORITHM:

The feature selection problem as a regression problem with a linear predictor. The extension of the elastic net regularize into a structured elastic net. The structured elastic net regularize is of the following form:

$$\alpha||\beta||_1 + (1-\alpha)\beta^T \Lambda \beta$$

where, $\Lambda[p\times p]$ is assumed to be symmetric and positive semi definite.β is a p-dimensional vector representing the regression coefficients. Take Λ as the correlation matrix, then the structured elastic net is the linear combination of lasso and redundancy removal term.

Assume a loss function to estimate the β values subject to the structured. Elastic net regularizer constraint. So, their optimization problem is defined as

$$\hat{\beta} = argmin_{\beta} \sum_{i=1}^{n} L(y_i, f(x_i, \beta))$$

Subject to

$$\alpha||\beta||_1 + (1-\alpha)\beta^T \Lambda \beta \le s,\ \alpha \in (0,1),\ s > 0$$

choose the feature fj for inclusion in S is as follows:

$$f_j = \arg \max_{fi} \{I(c; f_i) - 1/|S| \sum_{f_s \in S} NI(f_i, f_s)$$

This process is repeated until the desired number of features is chosen. NI(f_i; f_s) is the normalized mutual information, which is defined as

$$NI(f_i; f_s) = I(f_i; f_s) / (\min\{H(f_i), H(f_s)\}).$$

decision dependent (supervised) and decision-independent (unsupervised) correlation-based feature selection (CFS) approaches. The decision-dependent correlation (DDC) between feature f_i and f_j given decision y is defined as

$$Q_y(f_i; f_j) = (I(y; f_i) + I(y; f_j) - I(y; f_i; f_s))/H(y)$$

where I and H are the mutual information and entropy, respectively.

ID3 algorithm is used to generate a decision tree. It is a precursor to the C4.5 Algorithm. Classifies data using the attributes. Tree consists of decision nodes and decision leafs. Nodes can have two or more branches which represents the value for the attribute tested. Leafs nodes produces a homogeneous result. The ID3 follows the Occam's razor principle. Attempts to create the smallest possible decision tree. Take all unused attributes and calculates their entropies. Chooses attribute that has the lowest entropy is minimum or when information gain is maximum. Makes a node containing that attribute.

$$\text{Entropy(S)} = \sum_{i=1}^{c} P_i \log_2 P_i$$

And with entropy values information gain is calculated. Initially information gain is calculated for each table considering the entropy values that are found. Finally the root node is established with the help of the information gain.

5. DESCRIPTION OF EXISTING AND PROPOSED SYSTEM:

Finding the optimal subset of features usually requires an exhaustive search considering all possible subsets of features, which becomes computationally prohibitive when the dimension of the data is high. Therefore, even for a wrapper method, some suboptimal heuristic guided selection methods are used. Use of forward selection or backward elimination schemes or their variants may not be able to exploit the interaction between features. The effectiveness of a feature set depends not only on the problem, but also on the tool that is used to solve the problem.

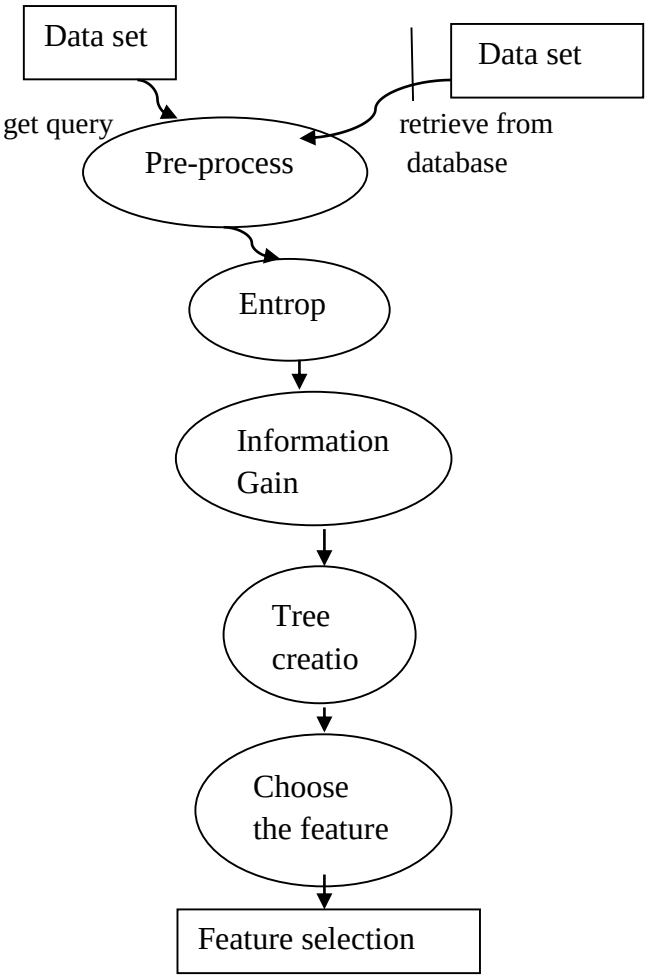

Fig.1 Dataflow Diagram

The best way of feature selection should be a combined approach where the learning system scans at all features concurrently and picks up valuable features while designing the system for evaluating the specified problem. There are a few methods that address the feature selection problem using such an integrated framework. This type of methods has some advantages over other techniques: only assess wanted feasible subsets not all subsets, they can report for relation between features, and they can recognize the interaction between the features and the tool that is used to get to the bottom of the problem. Various algorithms and techniques like categorization (classification), grouping, degeneration, nearest neighbor method, artificial intelligence, neural networks, decision trees, genetic algorithm, relationship rules, etc., are used for knowledge discovery from databases.ID3 Algorithm is put into operation to offer better results. The projected scheme, named as Feature Selection with controlled redundancy (FSMLP-CoR) is to be implemented.

A large dataset of university is taken. Data pre-processing includes cleaning, standardization, modification, attribute extraction and selection. Entropy is a measure of impurity (the opposite). It is defined for a binary class with values a and b as follows:

Entropy = - p(a) * log(p(a) – p(b) * log(p(b))

Decision tree builds classification or degeneration models in the form of a tree structure. It splits the dataset into smaller and smaller subsets while at the same time a related decision tree is progressively developed.

6. EXPERIMENTAL RESULTS:

Modulator and activation functions were used in the current work. The initial values of the modulator functions are set such that the gates are almost closed, i.e., the values are almost zero. Discard an attribute (feature) if its shrinking value is more than 90.0%. For a specified problem, one can discover an suitable threshold using a cross-validation technique.

Every data set is standardized using the formula:

$$x=(x - \mu)/\sigma$$

Where, x is the standardized feature value, μ and σ are the mean and standard deviation of the attribute. Most of the experiments with FSMLP-CoR, use just one unseen layer. To make results more consistent, use two level cross-validation techniques. In the outer level, initially, partition the data randomly into ten folds of similar size. After the attributes are chosen, project Y on the selected feature space and call the projected version of Y as Y.

1.To assess how good these selected features are, train a conventional MLP using the selected features, i.e., using the data set Y.

2.To find the most desirable architecture, n2 vary the number of hidden nodes from two to eight. Next, train

INTERNATIONAL JOURNAL FOR INNOVATIVE RESEARCH IN MULTIDISCIPLINARY FIELD ISSN: 2455-0620 Volume - 6, Issue - 4, Apr – 2020
Monthly, Peer-Reviewed, Refereed, Indexed Journal with IC Value: 86.87 Impact Factor: 6.497
Received Date: 13/04/2020 Acceptance Date: 28/04/2020 Publication Date: 30/04/2020

an MLP with n2 hidden nodes using data set Y and test it on Xj, where Xj is the projected version of Xj that was left out in the outer loop. This process is repeated for all Xj; j =1, ...,10 in the outer loop to get the misclassification rate using the selected features.

3.Finally, the entire process is repeated 30 times, every time using a different random partition in the outer loop. The training is terminated when either the misclassification error reduces to less than 10% or the number of iterations reaches 1000.

Then makes a comparison of our two learning schemes with RCFS, which also selects features avoiding redundancy. RCFS cannot be used for regression/function approximation type problems. RCFS needs the user to specify the desired number of features. Using the results of previous outcomes find the smallest λ for which the maximum correlation between the pairs of selected features is less than 0.70. For a given data set, our method selects k features using RCSF. To realize RCFS, use the single linkage hierarchical clustering algorithm. The best results are in bold face.

To measure the classification accuracy, an MLP network with one hidden layer is used. For the very simple synthetic data, the maximum correlation remains the same for all the three methods. These results demonstrate that these methods are more effective in controlling redundancy compared with a state-of-the-art method. It exhibits a significant improvement in accuracy. Even for FSMLP-CoR, the test accuracy is better for 15 data sets compared with that by RCFS. It improves generalization by these algorithms. The final list of selected features was determined by considering the performance of an SVM classifier of RCFS. Since an MLP is used as the classifier, to make a fair comparison, MLP network is used to determine the performance of RCFS selected features.

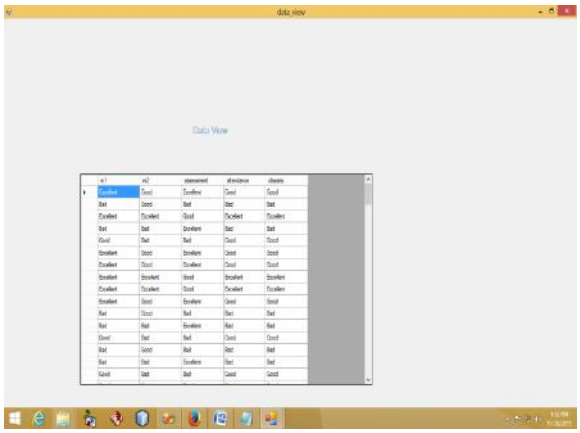

Fig.2 Database View

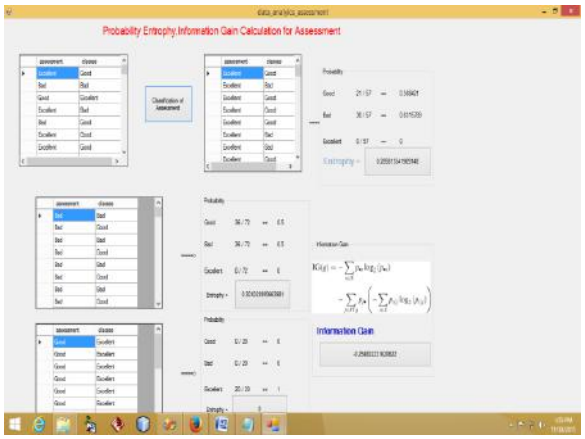

Fig.3 Information Gain

Fig.4 Tree Generation

7. CONCLUSION AND FUTUREWORK:

The feature selection using neural schema was implemented to find out the quality feature based on the dataset. With the help of the ID3 Algorithm the iterative structure is developed and the large dataset is pruned to remove the duplicates and the tree is created with the calculation of the entropy, information gain. The maximum information gain in the iterative layers is assigned as the root node and the sub-nodes are assigned by repeating the ID3 algorithm. The output is framed by checking the value from the tree node.

The novelty of the method is that it is an integrated scheme, which looks at all the features together while designing the decision-making system, and in this process, it picks up the features needed to solve the problem with a control on the level of

redundancy in the selected features. Thus, it does not require to evaluate different subsets of features. This system can account for the subtle interaction between features. It can also account for interaction between features, tools, and the problem being solved. The framework is very general and can be easily adapted to other networks and learning systems. In the future work, the database size will be increased and duplicate records are also can be identified to reduce the access time of the nodes. The efficiency can be further improved to increase the quality of the feature selection data.

REFERENCES:

1. M.S.Vijaykumar, "Fuzzy score based short text understanding from corpus data using semantic discovery", International Journal for Research & Development in Technology , ISSN: 2349-3585, Volume-9,Issue-3,2018.

2. L.Zhou, L.Wang, and C.Shen(May.2012), "Feature selection with redundancy constrained class separability" IEEE Trans. Neural Netw., vol.21, no.5, pp.853–858.

3. P.A.Estévez, M.Tesmer, C.A.Perez, and J.M.Zurada(Feb.2010), "Normalized mutual information feature selection," IEEE Trans. Neural Netw., vol.20, no.2, pp.189–201.

4. R.Liu, N.Yang, X.Ding, and L.Ma (Nov.2009), "An unsupervised feature selection algorithm: Laplacian score combined with distance-based entropy measure," in Proc. 3rd Int.Symp.Intell.Inf.Technol.Appl., pp.65–68.

5. C.Boutsidis, M.W.Mahoney, and P.Drineas(2009), "Unsupervised feature selection for the k-means clustering problem," in Proc. NIPS, , pp.153–161.

6. M.W.Mahoney and P. Drineas(2009), " CUR matrix decompositions for improved data analysis," Proc. Nat. Acad. Sci., vol.106, no.3, pp.697–702.

7. C.Shen, H.Li, and M.J.Brooks(2008), "Supervised dimensionality reduction via sequential semidefinite programming," Pattern Recognit., vol. 41, no. 12, pp. 3644–3652.

8. H.Liu and L.Yu(Apr.2005), "Toward integrating feature selection algorithms for classification and clustering," IEEE Trans. Knowl. Data Eng., vol. 17, no. 4, pp. 491–502.

9. F.Nie, S.Xiang, Y.Jia, C.Zhang, and S.Yan(Jul.2008), "Trace ratio criterion for feature selection," in Proc. 23rd AAAI Conf. Artif. Intell., pp. 671–676.

10. L.Wang(Sep.2008), " Feature selection with kernel class separability," IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 9, pp. 1534–1546.