International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A Novel Methodology for Feature Subset Selection using TLBO Algorithm

Rajeev[1], Dr. Rajdev Tiwari[2]

[1]Student UPTU, Noida Institute of Engineering & Technology, Greater Noida, India

[2]Associate Professor, Noida Institute of Engineering & Technology, Greater Noida, India

Abstract: ***In the present paper, a novel method for Feature Subset Selection in dataset, FSS-TLBOA (Feature Subset Selection by Teaching Learning Based Optimization Algorithm), is proposed. A dataset can contain several features. Many Clustering methods are designed for clustering low–dimensional data. In high dimensional space finding clusters of data objects is challenging due to the curse of dimensionality. When the dimensionality increases, data in the irrelevant dimensions may produce much noise and mask the real clusters to be discovered. To deal with these problems, an efficient feature subset selection technique for high dimensional data has been proposed. Feature subset selection reduces the data size by removing irrelevant or redundant attributes. Experiments are performed on the bank dataset to classify, according to the 11 existing features, with the help of TLBO (teacher learning based optimization) algorithm. This paper describes the main idea of Feature Subset Selection, presenting related work about each concept. Its aim is to improve the performance results of classifiers but using a significantly reduced set of features.***

Keywords: Features selection, Teaching-learning-based optimization algorithm.

1. Introduction

A data warehouse is designed to unify and maintain all the attributes that are relevant for the analysis processes. Due to the fast increasing in the size of the new operational systems, it neither necessary, nor becomes practical to load and maintain every operational attributes in the data warehouse.

Data sets consist of a number of vectors, each corresponding to some occurrence of an event: each vector consists of a large number of attributes. Many of these features are irrelevant or redundant. Unnecessary features increase the size of the search space and make generalization more difficult to capture rules necessary for forecasting or classification, whether by machine or by hand.

Data reduction is the process of minimizing the amount of data that requires to be stored in a data storage environment and which can increase storage efficiency and reduce costs. It can be achieved by using various different types of techniques. These are Data cube aggregation, Dimensionality reduction / attribute subset selection for removes unimportant attributes.

We are focusing especially to dimensionality reduction problem our work, which has been solved by various statistical and evolutionary methods in the past, but in this work we are use Attributes subset selection technique to solve this problem. Attributes subset selection as a method of dimensionality reduction is a well known optimization problem. We present a novel methodology for automated subset selection of the most relevant independent feature in a data warehouse. The feature subset selection method is based on the TLBO (Teaching Learning Based Optimization) approach to knowledge discovery in data warehouses.

1.1 Preprocessing of Data

In Data warehouse a number of different tools and methods used for data pre-processing. Data integration blends data into a coherent data store from multiple sources, as in data warehousing. Data cleaning routines attempts to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. These sources may be including multiple data cubes, databases, or flat files. In data transformation, the data are consolidated or transformed into the forms appropriate for mining. Data transformation can involve the Aggregation, Smoothing, and Normalization. Data reduction technique can be implemented to obtain a reduced representation of the data set that is too much smaller in volume, yet intimately maintains the integrity of the original data set. Data reduction method can reduce the data size by Aggregation, Sampling, Dimensionality Reduction, Feature subset selection, Feature creation and Attribute transformation, for instance.

1.2 Feature Subset Selection

Feature subset selection (FSS) is a search problem for finding an optimal or suboptimal subset of m features out of original M features. With the aim of selecting a subset of best features with respect to the target concepts, Feature subset selection is an effectual way for reducing data dimensionality, eliminating irrelevant data and redundant data, increasing accuracy and improving result comprehensibility.

FSS is one of the key types of data reduction. The main objective of this step is to find useful features that represent the data and remove those features that are either irrelevant or redundant. This is because an irrelevant feature doesn't provide any useful information to predict the target concept and redundant feature doesn't add extra information that might be useful to predict the target concept.

Feature subset selection involves in identifying a best subset of useful features that produces compatible results as the original entire set of features. This is since redundant features do not redound to getting a better predictor and irrelevant features do not contribute to the predictive accuracy for that they provide mostly information which is already present in other feature(s).

FSS helps in a number of ways e.g. it reduces useless features to save computing time and data storage, relevant features improves predictive performance and precludes over-fitting, provides more appropriate description of the target concept. Feature subset Selection is a combinatorial optimization problem where a feature set containing N number of features can be too large.

A number of feature subset selection methods have been studied and proposed for machine learning applications. These can be divided into four categories: the Wrapper, Filter, Embedded, and Hybrid approaches.

- **Wrapper methods**: The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets, the accuracy of the learning algorithms is usually high. However, the generality of the selected features is limited and the computational complexity is large.
- **Filter method**s: The filter methods are independent of learning algorithms, with good generality. Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed.
- **Embedded methods:** The embedded methods incorporate feature selection as a part of the training process and are usually specific to given learning algorithms, and therefore may be more efficient than the other three categories. Traditional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches.
- **Hybrid methods:** The hybrid methods are a combination of wrapper and filter methods, by using a filter method to reduce search space that will be considered by the subsequent wrapper. The primarily focus on combining wrapper and filter methods to achieve the best possible performance with a particular learning algorithm with similar time complexity of the filter methods
-

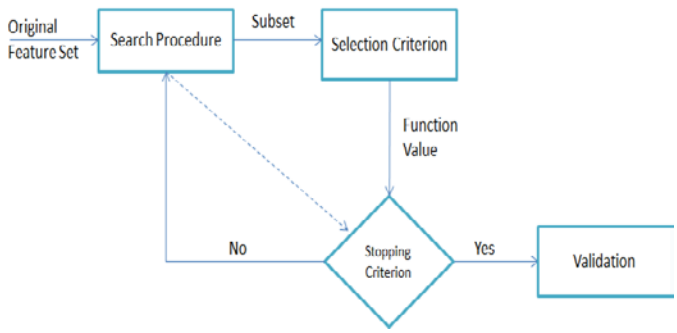

Figure 1: Four steps of feature selection

1.3 TLBO and Feature Subset Selection

Teaching-learning-based optimization algorithm (TLBO) is a teaching-learning process which is based on the effect of influence of a teacher on the output of learners in a class. TLBO is population based method, in this optimization algorithm a group of learners is considered as population and different design variables are considered as different subjects offered to the learners and learners' result is analogous to the 'fitness' value of the optimization problem. In the entire population the best solution is considered as the teacher.

The main idea behind TLBO is the simulation of a classical school learning process that consists of two stages. Teacher and learners are the two vital components of this algorithm and describes two basic modes of the learning, through teacher (known as teacher phase) and interacting with the other learners (known as learner phase).

During the teacher stage, a teacher imparts knowledge directly to his/her students. The better the teacher, the more knowledge the students obtain. Most students will partially accept new learning and, in some cases, the teacher will have almost no direct effect on students' knowledge. However, the possibility for most students to obtain new knowledge is not completely lost.

During the learner stage, a student may learn with the help of fellow students. Overall, how much knowledge is transferred to a student does not only depend on his/her teacher but also on interactions amongst students through peer learning. The output in TLBO algorithm is considered in terms of grades or results of all learners which depend on the quality of teacher. So, teacher is usually considered as a highly learned person who trains learners so that they can have better results in terms of their marks or grades. Moreover, learners also learn from the interaction among themselves which also helps in improving their results.

For feature subset selection, Teaching-learning-based optimization is used as a optimization selection algorithm.

2. Related Work

In 1997 Ron Kohavi, George H. John uses the wrapper method searches for an optimal feature subset tailored to a particular algorithm and a domain. He compares the wrapper approach to induction without feature subset selection and to Relief, a filter approach to feature subset selection. For the evaluation function, he used cross-validation as the accuracy estimation technique [1]. On the other hand in 2000 I. Inza, P. Larranaga, R. Etxeberria, B. Sierra presented FSS-EBNA (Feature Subset Selection by Estimation of Bayesian Network Algorithm), a new method for Feature Subset Selection in machine learning. FSS-EBNA is an evolutionary, randomized search algorithm, population-based, and it can be executed when domain knowledge is not available. FSS-EBNA, based on the Estimation of Distribution Algorithm paradigm, avoids the use of crossover and mutation operators to evolve the populations, in contrast to Genetic Algorithms [2].

In 2004 Lei Yu, Huan Liu is proposed a new framework of efficient feature selection via relevance and redundancy analysis, and a correlation-based method which uses *C*-correlation for relevance analysis and both *C*- and *F*-correlations for redundancy analysis and comparing the results with three representative feature selection algorithms [3]. For the effectiveness of feature selection a new approach Scatter Search metaheuristic is introduced by Felix Garcıa Lopez in 2004. It is the combination of Greedy Combination and the Reduced Greedy Combination. These methods provide two sequential algorithms that are compared with a recent Genetic Algorithm and with a parallelization of the Scatter Search. This parallelization is obtained by running simultaneously the two combination methods. Parallel Scatter Search presents better performance than the sequential algorithms [4].

For a better features subset selection a hybrid algorithm, SAGA is presented by Iffat A.Gheyas, LeslieS.Smith in 2009 for Searching for an optimal feature subset from a high dimensional feature space is known to be an NP- complete problem. SAGA combines the ability to avoid being trapped in a local minimum of simulated annealing with the very high rate of convergence of the crossover operator of genetic algorithms, the strong local search ability of greedy algorithms and the high computational efficiency of generalized regression neural networks. they compare proposed algorithm against the following benchmark algorithms: GA, PSO, SBS, SFBS, ACO, FW, SA, SFFS and SFS on both synthetic and real-world datasets. Among these datasets, one dataset has 285 features and the remaining 29 datasets have 10,000 features each. They study the performance of these algorithms at different time intervals: after 8, 16, 24, 72, 168 and 240 h of running and the performance of their algorithm is highly encouraging [7].

Houtao Deng, George Runger proposed a tree regularization framework, which adds a feature selection capability to many tree models in 2012. He applied the regularization framework on random forest and boosted trees to generate regularized versions (RRF and RBoost, respectively). After this Experimental studies show that RRF and RBoost produce high-quality feature subsets for both strong and weak classifiers. As tree models are computationally fast and can naturally deal with categorical and numerical variables, missing values, different scales (units) between variables, interactions and nonlinearities etc., the tree regularization framework provides an effective and efficient feature selection solution for many practical problems [12].

3. Proposed Method

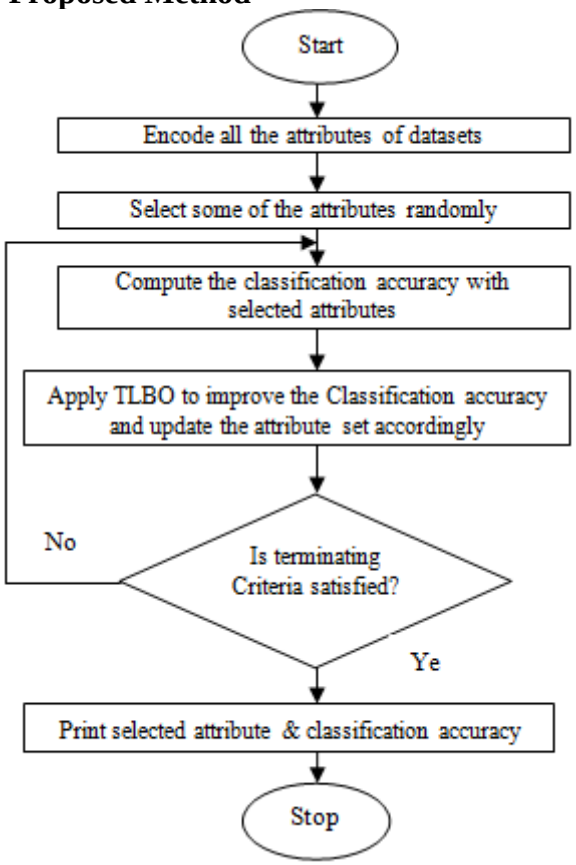

Figure 2: Proposed method for selecting optimal subset of attributes

Apply TLBO approach to Optimization of dataset and classification accuracy by using farther following steps:
Step 1:
Define the optimization problem as Minimize or Maximize f(X)
Where f(X) is the objective function value and X is a vector for design variables.
Step 2:
Initialize the population (i.e. learners, k=1,2,…,n) and design variables of the optimization problem
(i.e., number of subjects offered to the learners, j=1,2,…,m), and evaluate them.
Step 3:
Select the best solution who acts as chief teacher for that cycle.
Step 4:
Select the other teachers (T) based on the chief teacher and rank them,
(If the equality is not met, select the f(X)s closest to the value calculated above)
Step 5:
Assign the learners to the teachers according to their fitness value as:
For k=1:(n-s)
If f(X)1=f(X)k>f(X)2,
assign the learner f(X)k to teacher 1 (i.e f(X)1).
Else, if f(X)2=f(X)k>f(X)3,
assign the learner f(X)k to teacher 2 (i.e f(X)2).
Else, if f(X)T-1=f(X)k>f(X)T,
assign the learner f(X)k to teacher ‘T-1’ (i.e f(X)T-1).

Else,
assign the learner f(X)k to teacher 'T'
End
Step 6:
Keep the elite solutions of each group.
Step 7:
Calculate the mean result of each group of learners in each subject.
Step 8:
For each group, evaluate the difference between the current mean and the corresponding result of the teacher of that group for each subject by utilizing the adaptive teaching factor.
Step 9:
For each group, update the learners' knowledge with the help of the teacher's knowledge, along with the knowledge acquired by the learners during the tutorial hours.
Step 10:
For each group, update the learners' knowledge by utilizing the knowledge of some other learners, as well as by self learning.
Step 11:
Replace the worst solution of each group with an elite solution.
Step 12:
Eliminate the duplicate solutions randomly.
Step 13:
Combine all the groups.
Step 14:Repeat the procedure from step 3 to 13 until the termination criterion is met.

Table 4.1: Result Comparison

Id	*Sex*	*Age*	*region*	*Income*	*Married*	*Car*	*Children*	*Save_acc*	*Current_acc*	*Mortgage*	*pep*
ID12101	FEMALE	48	INNER_CITY	17546	NO	NO	1	NO	NO	NO	YES
ID12102	MALE	40	TOWN	30085.1	YES	YES	3	NO	YES	YES	NO
ID12103	FEMALE	51	INNER_CITY	16575.4	YES	YES	0	YES	YES	NO	NO
ID12104	FEMALE	23	TOWN	20375.4	YES	NO	3	NO	YES	NO	NO
ID12105	FEMALE	57	RURAL	50576.3	YES	NO	0	YES	NO	NO	NO
ID12106	FEMALE	57	TOWN	37869.6	YES	NO	2	YES	YES	NO	YES
ID12107	MALE	22	RURAL	8877.07	NO	NO	0	NO	YES	NO	YES
ID12108	FEMALE	58	TOWN	24946.6	YES	YES	0	YES	NO	NO	NO
ID12109	MALE	37	SUBURBAN	25304.3	YES	YES	2	NO	NO	NO	NO
ID121010	FEMALE	54	TOWN	24212.1	YES	YES	2	YES	YES	NO	NO
ID121011	FEMALE	66	TOWN	59803.9	YES	NO	0	YES	YES	NO	NO
ID121012	FEMALE	52	INNER_CITY	26658.8	NO	YES	0	YES	YES	YES	NO
ID121013	FEMALE	44	TOWN	15735.8	YES	NO	1	YES	YES	YES	YES
ID121014	FEMALE	66	TOWN	55204.7	YES	YES	1	YES	YES	YES	YES

4. Implementation and Result Discussion

The optimal subset of attribute obtained on executing the proposed algorithm is then tested for its classification accuracy using SIPINA tool of the TANAGRA Software. This is free data mining software which used for research and academic purposes. By proposed method, we get better classification accuracy in comparison to wrapper based and filter based methods. In table 4.1, a comparison result of the proposed method to other method in terms of classification accuracy is shown.

Table 4.1: Result Comparison

Sr. No	*Methods*	*Attributed subset selected*	*% Classification accuracy*
	Wrapper based methods		
1	forward Selection Multicross Validation	(Children, age, married, mortgage, region)	72.33
	Bootstrap backward elimination	(Sex, region, income, married, children, car, save_acc, current_acc)	73.33
	Filter based methods		
3	Relief	(Children, Save_acc)	68.50
4	MIFS	(Children, Married, Income, Sex)	72.66
	Proposed method		
5	TLBO	(Children, Married, mortgage, sex, income, save_acc)	74.07

5. Conclusion

This paper presented a new feature subset selection algorithm for high dimensional data in the data ware house. A number of feature selection methods are compared with the proposed method. It performs either equally good or better than many of the other existing methods. The capabilities of the proposed method have shown in previously results. Its accuracy is more when applied on large and real dataset.

References

[1] Ron Kohavi and George H. John. 1997. Wrappers for feature subset selection. Artificial Intelligence 273-324
[2] Inza, P. Larranaga, R. Etxeberria, B. Sierra. 2000. Feature Subset Selection by Bayesian network-based optimization. Artificial Intelligence 123 157–184
[3] Isabelle Guyon and Andr´e Elisseeff. 2003. An Introduction to Variable and Feature Selection. Journal of Machine Learning Research 3 1157-1182
[4] Lei Yu and Huan Liu. 2004. Efficient Feature Selection via Analysis of Relevance and Redundancy. Journal of Machine Learning Research 5 1205–1224
[5] Fe´lix Garcı´a Lo´pez. 2004. Miguel Garcı´a Torres, Bele´n Melia´n Batista, Jose´ A. Moreno Pe´rez *, J. Marcos Moreno-Vega Solving feature subset selection problem by a Parallel Scatter Search. European Journal of Operational Research 169 (2006) 477–489
[6] Dunja Mladeni. 2006. Feature Selection for Dimensionality Reduction SLSFS 2005, LNCS 3940, pp. 84–102
[7] Iffat A.Gheyas and LeslieS.Smith. 2010 Feature subset selection in large dimensionality domains. Pattern Recognition 43, 5 -13
[8] Rajdev Tiwari and Manu Pratap Singh. 2010. Correlation-based Attribute Selection using Genetic Algorithm International Journal of Computer Applications (0975 – 8887)Volume 4– No.8
[9] Yijun Sun, Sinisa Todorovic, and Steve Goodison. 2010. Local Learning Based Feature Selection for High Dimensional Data Analysis IEEE transactions on pattern analysis and machine intelligence, vol. 32, no. 9
[10] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset. Expert Systems with Applications 38 5947–5957
[11] Asha Gowda Karegowda M.A.Jayaram. 2011. A.S .Manjunath Feature Subset Selection using Cascaded GA & CFS: A Filter Approach in Supervised Learning. International Journal of Computer Applications (0975 – 8887) Volume 23– No.2
[12] Yuanning Liu1,2, Gang Wang, Huiling Chen, Hao Dong1,2, Xiaodong Zhu, Sujing Wang. 2011 An Improved Particle Swarm Optimization for Feature Selection. Journal of Bionic Engineering 8 191–200
[13] Houtao Deng and George Runger. 2012. Feature Selection via Regularized Trees
[14] Syed Imran Ali and Waseem Shahzad. 2012. A Feature Subset Selection Method based on Symmetric Uncertainty and Ant Colony Optimization. International Journal of Computer Applications (0975 – 8887) Volume 60– No.11
[15] Qinbao Song, Jingjie Ni and Guangtao Wang. 2012. A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data. ieee transactions on knowledge and data engineering vol:25 no 1

Author Profile

Rajeev received the B.Tech degree in Information technology from the Noida Institute of Engineering Technology affiliation to UPTU, Lucknow in 2011, and Pursuing M.Tech degrees in Computer Engineering from Noida Institute of Engineering Technology. His research interest area is Data mining.

Dr. Rajdev Tiwari (Director MCA, Noida Institute of Engineering Technology, Gr.Noida) is a mult-faceted personality with rich corporate, government and academic experience. He is qualified UGC NET (Computer Science), Ph.D. (Computer Science), MCA, Post Graduate Diploma in Advance Software Design & Development, M Sc and B Sc.