Adaptive hybrid methods for Feature selection based on Aggregation of Information gain and Clustering methods

P. Ranjit Jeba Thangaiah†, R. Shriram††, and K. Vivekanandan†††

†Research Scholar, Department of Computer Science and Engineering, Bharathiar University, Coimbatore, India
††Assistant Professor, B.S. Abdur Rahman University, Chennai, India
†††Reader, BSMED, Bharathiar University, Coimbatore, India

Summary

The growing abundance of information necessitates the need for appropriate methods for organization and evaluation. Mining data for information and extracting conclusions has been a fertile field of research. However data mining needs methods to preprocess the data. Feature selection is a growing field of interest about selecting proper information from information repositories. The aim of this paper is to highlight the need for feature selection methods in data mining encompassing the best characteristics of the data. In recent times there has been interest in developing hybrid feature selection methods combining the characteristics of various filter and wrapper methods. The proposed method advocates an adaptive aggregation strategy using a) the gain ratio for candidate features and b) clustering methods to find the distribution of candidate features. The underlying principle of the strategy is that the best individual features need not constitute the best sub-set of features representing the problem. A given feature might provide more information when present with certain other feature(s) than when considered by itself. The adaptive method has been implemented for the datasets from the UCI repository and results correlated. The conclusions show that the proposed method shows encouraging results.

Key words:
Data mining, Feature selection, Adaptive method, Correlation Clustering.

1. Introduction

Data mining is the process of discovering non-trivial, previously unknown and potentially useful information from large volumes of data [13]. Though not a new activity, it is becoming more popular as the scale of data has increased due to the advent of the data processing, computing technology and Internet. Data mining can forecast future trends and activities to support of people [11]. Data preprocessing is very important to successful data mining [9] techniques. Feature selection is a term space reduction method which attempts to select the more discriminative features from data sets in order to improve classification quality and reduce computational complexity [17].

Feature selection methods can work on labeled and unlabeled data [6]. A goal of feature selection is to avoid selecting too many or too few features than is necessary. If too few features are selected, there is a good chance that the information content in this set of features is low. On the other hand, if too many (irrelevant) features are selected, the effects due to noise present in (most real-world) data may overshadow the information present. Hence, this is a trade-off that must be addressed by any feature selection method.

There are three fundamentally different approaches for feature selection: wrapper, filter and hybrid. Wrapper approach uses the learning algorithm to test all existing features subsets. The filter approach corresponds to a data pre-processing step preceding the learning phase. The wrapper algorithms use the actual classifier to select the optimal feature subset, while the filter features independently of the classifier [10]. The fundamental difference between the two families is that the wrapper algorithms are related to the learning algorithm whereas the filter phase is completely independent of it.

The disadvantage of filter approach is that the features could be correlated among themselves [2]. On the other hand, wrapper methods tend to find features better suited to the predetermined learning algorithm resulting in better performance. But, it also tends to be more computationally expensive since the classifier must be trained for each candidate subset [1]. It is also possible to combine the filter and wrapper methods to obtain hybrid approaches [16]. In recent literature, the term embedded methods [8] has been introduced. The inference system has its own Feature Selection Algorithm (either explicit or implicit) in embedded methods.

Manuscript received February 5, 2009
Manuscript revised February 20, 2009

The rest of the paper is organized as follows. In section 2, the comparison of relevant literature is presented to establish the state of art and outline the basis for the uniqueness of this work. Section 3 describes the various aspects of the proposed method. The implementation of the system and the relevant results are given in section 4. Section 5 concludes the paper.

2. Literature Review

Feature subset-selection focuses on the reduction of unnecessary features or objects in a given data set to improve the performance of some algorithm that is trying to solve a problem. Whatever be the method, the key aspect in feature space is the search heuristic as feature selection is fundamentally a process of selecting a subset of features from a power set of features. There are three kinds of heuristic search methods [15] proposed in literature: forward selection, backward elimination and genetic algorithms.

A hybrid two stage feature selection algorithm has been proposed by Vidyavathi and Ravikumar [14]. The method uses the Fischer's gain ratio and mutual information. The proposed approach goes beyond this work in that it brings out the multi-dimensionality of the subset and works equally well for small and large data sets. Michalak and Kwasnicka [7] have advocated a correlation based feature selection strategy to take into account the interdependency between variables. The proposed approach has incorporated clustering as a higher level mechanism for correlation. But, correlation can be included in the work in future. The proposed method also assimilates the work of Zhang, Chen and Zhou [3] who have advocated an approach involving pair-wise constraints using a filter method.

Ienco and Meo [4] have proposed a hierarchical clustering algorithm based on distance measures. Distance metrics used in the proposed work are similar to this approach. The difference in the proposed approach is that clustering is incorporated as one aspect in overall method of feature selection. The proposed feature selection method builds on the method used by She et al. [12] where a two-stage method is described. The proposed method here has used a clustering and information gain calculation approach. Also the work of Legrand and Nicoloyannis [5] is an important reference for this work. The algorithm uses a three-stage approach aggregating the information from ten different methods and three metrics and validating the algorithm accordingly. The proposed work uses a five-stage approach in which the clustering approach supplemented with information gain based method incorporates a learning algorithm for the processing. Thus, the proposed work is a hybrid of clustering, information gain, correlation and learning algorithms. While each approach by itself has generated appreciable results, it is perhaps time to use a logical hybrid especially for large data sets with more number of features. In the next section the proposed method is described.

3. Proposed method

This work describes a hybrid method for selecting relevant features in classification problems. This approach does not belong to wrapper approach or filter approach. The major difference is that the approach is used by the data mining system to learn information and heuristics for the larger classification task. Thus the information can be used to adapt and react appropriately to newer data values without necessitating the larger analyse-classify cycle again.

The model (Figure 1) analyses a data set, uses a repository of hybrid techniques for feature selection model based technique. The measures used for this purpose are

a) Information Gain: The parametric importance of a feature in a data set is found by using this measure. There are many methods for information gain calculation.
b) Separability measure: The distance between each candidate feature and other candidate features are found using this value. The distance measures are found using conditional probability values that can separate the classes as far as possible.
c) Class relationship: This gives a measure of the degree of membership of each feature in each class in the data set.
d) Class distance: This value gives the distance of each class from others.

The proposed feature selection repository has five distinct steps.

3.1. Clustering:

This step scans the data set and identifies at a high level the various classes that are present in a data set. In this step the feature selection system searches the data set for various classes. The data set is modeled as a set of points by using pre-processing methods. Select any two variables at random. Calculate the distance metrics between them. Keep doing this all the variables have been selected as a part of some pair. Order the distances in ascending order. Now classify the variable pair with lowest distance metric as class 1. Now calculate the distances of class 1 with other variable pairs by using normalized threshold memberships.

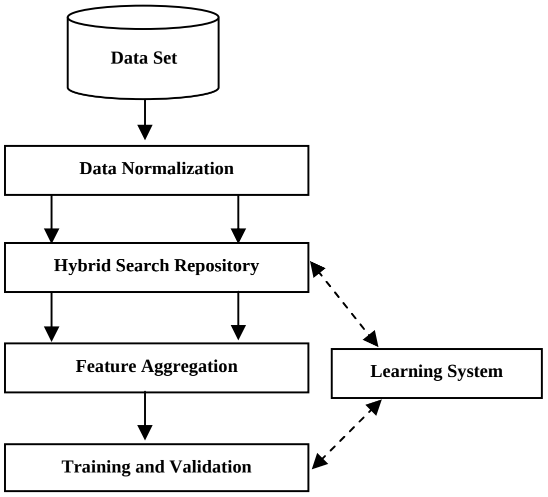

Fig.1 Feature Selection Model

This way keep building classes 1,2..n where n depends on the thesholding criterion. The aim is to find Cl_i such that it covers the data set. A feature (f_i) may contribute in different ways to any number of classes. The clustering system may not be able distinctly identify any classes at all in the system or find interlinked classes as well. The clustering approach helps in identifying the class relationship (Cr_{ij}) and the separability measure (S_{ij}) for a feature.

Thus the values found not only represent the values for single data set, but also help the system discover previously unknown relationships between various features due to the clustering measures. This will help classify hitherto unseen data items in case of new data and keep the predictor from developing a bias. For typical relevance measures, all information required in the computation is the class distribution associated with each feature. If the class information can be propagated to all data sources, it is easy to evaluate all features locally. Thus, instead of evaluating all features in a central table, the class labels are pushed into each individual data source.

3.2 Identifying the values for each measure (wrapper approach):

In this stage, the information gain measures are calculated using the Shannon's entropy method. While the information gain for each feature will give the measure of importance of each feature in isolation, the separability measure and Class relationship give a broad measure of the inter-relationships between the features and bring out the importance. Each criterion list is calculated in parallel and the feature lists are calculated.

3.3 Learning algorithm (filter approach):

The learning algorithm controls the learning of relationships between the features and the classes. The features subset identification systems data relationships are codified in the form of heuristic methods. Automated machine learning systems are used to learn explicit and implicit relationships between features and data sets. The learning algorithm changes its rules during the validation stage automatically. Data sets used in the feature classification sometimes are a poor representation of the incoming data and may bias the algorithms. To prevent that there is a need for an incremental learning strategy that combines the best of exhaustive search and heuristic search but also learning the characteristics of the data sets considered. If the accuracy of the learning algorithm is good, the performance of the overall system will be improved. This learning system then guides the validation process by not only learning from the validation process, but suggesting new feature classes if there are new classes of data that make some hitherto un-considered features relevant.

3.4 Features subset Aggregation (wrapper approach):

This step aggregates the data obtained in the previous steps and identifies the candidate measures incrementally. The advantage of having an incremental and exhaustive search is that it allows the feature selection method to iterate in a thorough manner without losing any aspect of the data set. *F* denotes a feature selection algorithm that selects the *n* most significant features from the original *d*-dimensional input space. Subspace projection is a randomized procedure that selects, according to the uniform distribution, a *k*-subset $A = (a_1, a_2, .. a_n)$ from (1,2,…n), so defining a projection PA : Rn ! Rk; where $PA(x1; : : : ; xn) = (xa1 ; : : : ; xak)$; then it returns as output the new k-dimensional data set $f(PA(xj); tj)$ *such that* $1<j < m$, where $D = (xj); tj$ such $1<j< m$ is the set of the *n*-dimensional features selected from the original *d*-dimensional input space. Every new data set *Di* obtained through the iteration of the procedure Subspace projection is given as input to a learning algorithm *L* which outputs a classifier *hi*. The learning algorithm selects the subset based on the distribution of the data set. If the dataset is highly clustered, then clustering is given equal importance as information gain. In case of a diverse and sparse data set where identifying clusters is very difficult, information gain based methods are given a higher importance than the clustering methods. Thus depending on the input distribution and previously known data, the obtained classifiers are finally aggregated through majority voting.

3.5 Validation of the optimal features subset:

The data set values are chosen at random to prevent generalization. The validation algorithm uses a randomly selected training set classified in advance. For this if a test case generator is available which will select data sets based on heuristic principles, it will be useful. Heuristic principles balance the data applied in terms of noisy and noise free connotations. These heuristics are needed as frequently training sets are based on a subset of data that does not represent the entire problem (loss of data during the pruning stages). These heuristic learning modes build on the active learning methods where the algorithms are used to choose which data set to apply next. The algorithm thus will become a semi-supervised model by which the unlabeled data are used to enhance the underlying technique.

4. Implementation

The aim is to bring out the important attributes in the dataset. Sometimes an attribute that seems to be important in human intelligence is found to be missing in the attribute selection techniques based on correlation. But there is a possibility in the ranking techniques that these attributes found to be important in the ranking order. Thus in this approach hybrid attribute selection procedure is carried out by aggregating the attribute subset from information gain, correlation, and clustering.

Seven Datasets from UCI Machine Learning Repository is taken to examine this approach ranging from 10 attributes to 10000 attributes. The results are tabulated and the results are shown in the graph. The adult dataset from UCI machine learning repository is clearly examined for different output to see that this approach work normal in giving the deliverable that other attribute selection technique gives.

In the first result, we aim to find the number of features selected by the various methods. This is dependent of course on the threshold parameters. In this paper, the threshold is calculated by the adaptive aggregation algorithm which selects the separation measure. This measure is the minimum number of features for separation. Beyond this measure, any more values will instead select the entire feature set. It is found that this hybrid approach gives more information (Figure 2) on the attributes that are missed in attribute selection technique based on correlation. The question next is whether the approach works equally well for data sets with large number of attributes.

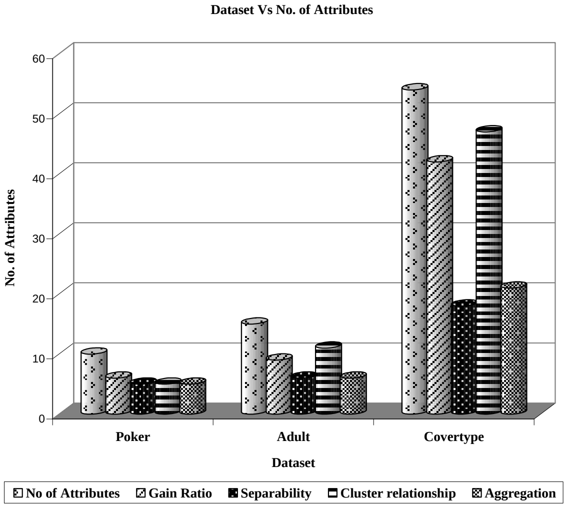

Fig. 2 Dataset with number of attributes less than 100

It is also found that even when the number of attributes is high the aggregation of attribute selection works (Figure 3) better in giving the number of necessary and important attributes because it takes account on correlation, ranking and distance based clustering technique. It is also observed that there are variations in the distribution characteristic in the methods like information gain, clustering and separability in different data sets whereas the proposed method works uniformly irrespective of the different data sets. This proves that the proposed approach is domain independent and works uniformly for data sets irrespective of the number of features.

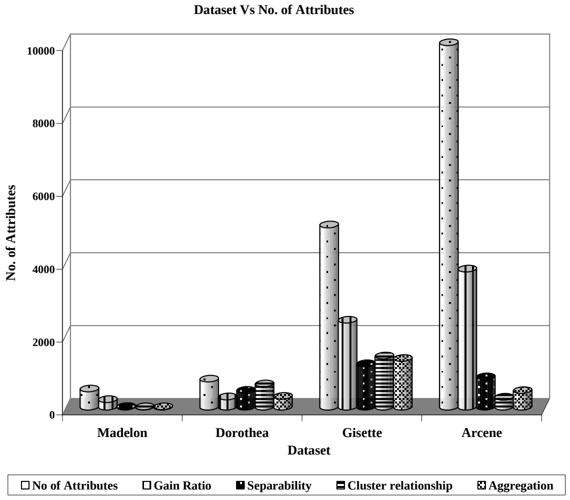

Fig 3 Dataset with number of attributes greater than 100

The pertinent question is whether these measures form an accurate measure of the overall data set. This issue is addressed by the validation aspect (Figure 4).

The proposed approach is applied on datasets with different number of attributes and characteristics and the resultant attribute subsets are checked for accuracy. The classification accuracy obtained by aggregation of attributes is given below. This shows that the feature subset selected by this proposed approach is well within the prescribed criteria of accuracy given in literature. This validates the conclusions on the algorithm.

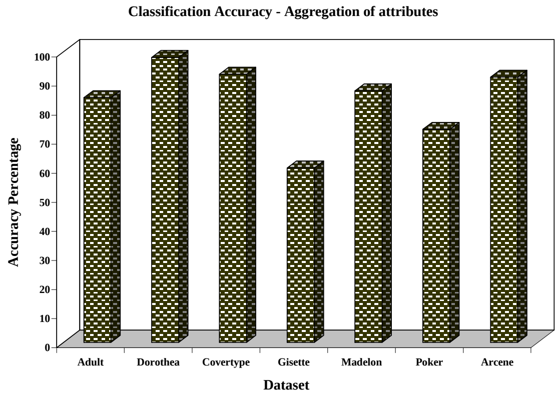

Fig. 4 Classification accuracy based on aggregation of attributes

5. Conclusion

Data Mining is not the answer to all problems and sometimes it has been over emphasized. It is expensive to carry out the entire process and therefore has to be thought out clearly. Feature selection approaches reduce the complexity of the overall process by allowing the data mining system to focus on what is really important. Thus, the data mining knowledge produced is found more meaningful. Also the new users / end users will get better results quickly. This research work validates a hybrid approach for feature selection. The results validate the initial objectives and are promising. Further work is envisaged in the approach incorporating correlation based approaches and expanding the methods for information gain.

References

[1] Agrawal. R. K. and Rajni Bala, (2007) “A Hybrid Approach for Selection of Relevant Features for Microarray Datasets”, International Journal of Computer and Information Science and Engineering Volume: 1, Number: 4, and Proceedings of World Academy of Science, Engineering and Technology Volume 23, ISSN 1307-6884, pp. 281 – 287.

[2] Chris Ding, Hanchuan Peng, (2003) "Minimum Redundancy Feature Selection from Microarray Gene Expression Data", IEEE Computer Society Bioinformatics Conference (CSB'03), pp.523.

[3] Daoqiang Zhang, Songcan Chen and Zhi-Hua Zhou, (2008) “Constraint Score: A New Filter Method for Feature Selection with Pairwise Constraints”, Pattern Recognition, Volume 41, Issue 5, pp. 1440 – 1451.

[4] Dino Ienco and Rosa Meo, (2008) “Exploration and Reduction of the Feature Space by Hierarchical Clustering”, Proceedings of the 2008 SIAM International Conference on Data Mining, Atlanta, Georgia, pp 577 – 587.

[5] Gaelle Legrand and Nicolas Nicoloyannis, (2005) “Feature Selection Method Using Preferences Aggregation”, Lecture Notes in Computer Science, Machine Learning and Data Mining in Pattern Recognition, Volume 3587/2005, Springer Berlin / Heidelberg.

[6] Huan Liu and Lei Yu, (2005) “Toward Integrating Feature Selection Algorithms for Classification and Clustering”, IEEE Transactions on Knowledge and Data Engineering, Volume 17, Issue 4, pp. 491 – 502. DOI:10.1109/TKDE.2005.66

[7] Krzysztof Michalak, Halina Kwasnicka, (2006) “Correlation–Based Feature Selection Strategy in Classification Problems”, International Jornal of Applied Mathematics and Computer Science, Volume 16, Number 4, pp. 503–511.

[8] Lal. T. N, Chapelle. O, Weston. J, and Elissee. A, (2006) “Embedded methods. In I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, editors, Feature Extraction: Foundations and Applications, Studies in Fuzziness and Soft Computing”, 207, pp. 137 - 165. Springer.

[9] Liu. H, and Motoda. H, (1998) “Feature Selection for Knowledge Discovery and Data Mining”, Boston: Kluwer Academic Publishers.

[10] Michal Haindl, Petr Somol, Dimitrios Ververidis, and Constantine Kotropoulos, (2006) “Feature Selection Based on Mutual Correlation”, Lecture Notes in Computer Science, Progress in Pattern Recognition, Image Analysis and Applications, Springer Berlin / Heidelberg, Volume 4225. pp. 569 – 577.

[11] Ni. X, (2008) “Research of Data Mining Based on Neural Networks”, Proceedings of World Academy of Science, Engineering and Technology, Volume 29.

[12] Rong She, Ke Wang, Yabo Xu, (2005) “Pushing Feature Selection ahead of Join”, SIAM International Conference on Data Mining

[13] Saravanan. V and Vivekanandan. K, (2004) “Design and implementation of automated data mining using intelligent agents in object oriented databases”, Intelligent information processing II, Springer-Verlag, pp. 221 - 226

[14] Vidyavathi, B. M, and Ravikumar. C. N, (2008) “A Novel Hybrid Filter Feature Selection Method for Data Mining”, Ubiquitous Computing and Communication Journal, Volume 3, Number 3, http://www.ubicc.org/files/pdf/Final%20manuscript-paper-ID237_237.pdf

[15] Yan Liu and John R. Kender, (2003) “Sort-Merge Feature Selection for Video Data”, Proceedings of the 2003 SIAM International Conference on Data Mining, San Francisco, CA, pp 321 – 325

[16] Yen-Po Lee, Wei-Yu Han, Wu-Ja Lin and Kuang-Shyr Wu, (2006) “A Hybrid Feature Selection Method Using Modular Perceptron Networks”, Asian Journal of Information Technology, Volume 5, Number 10, pp. 1088 – 1094

[17] Zhihua Wei, Jean-Hugues Chauchat and Duoqian Miao, (2008) “Comparing different text representation and feature selection methods on Chinese text classification using Character n-grams”, JADT 2008: 9th Journées internationales d’Analyse statistique des Données Textuelles, pp. 1175 – 1186.

P. Ranjit Jeba Thangaiah received the B.Sc. degree in Physics and M.C.A. degree from Bharathiar University in 1999 and 2002 respectively. He is with Karunya University but currently pursuing his research in the area of Data Mining at Bharathiar University.

K. Vivekanandan obtained his PhD in Computer Science from Bharathiar University during 1996. He has completed his post graduate in Applied Mathematics in 1981. He is with Bharathiar University since 1986. His research area includes Data Mining and Image Processing.

R. Shriram is an Assistant Professor at BS Abdur Rahman University, Chennai. He has done considerable work in Pervasive Computing, Mobile Computing and Language Computing. His research interests span mobile community networks and cross lingual information retrieval systems.