Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 5, May 2014, pg.480 – 485

RESEARCH ARTICLE

FEATURE SELECTION USING AN EFFECTIVE DIMENSIONALITY REDUCTION TECHNIQUE

Dr. M.V.Siva Prasad, CH. Suresh Kumar, T. Maneesha
Dept. of CSE, Anurag Engineering College, Kodad, Andhra Pradesh, India
principal@anurag.ac.in; suresh.mtech11@gmail.com; maneesha522@gmail.com

Abstract— processing applications with a large number of dimensions has been a challenge to the KDD (Knowledge Discovery and Data mining) community. An effective dimensionality reduction technique is an essential pre-processing method to remove noisy features. The proposed combined method for feature selection, where a filter based on correlation is applied on whole features set to find the relevant ones, and then, on these features a wrapper is applied in order to find the best features subset for a specified predictor.

Keywords— Feature subset selection; feature clustering; filters; wrappers

I. INTRODUCTION

The feature extraction methods such as Principal Components Analysis, Karhunen-Loeve or Singular Value Decomposition transformation, Pattern Classification and pattern recognition are the earliest methods for reducing dimensionality for unsupervised learning. These methods do not reduce the number of the original features; instead they create principal components or extracted features from the original features. Numbers of methods for feature selection for clustering are proposed in the last several years most of which are 'wrapper' in approach. In clustering, a wrapper method uses a clustering algorithm to evaluate the candidate feature subsets. Wrapper methods can be categorized into global type and local type. The global type assumes a subset of features to be more important than others for the whole data while the local type assumes each cluster to have a subset of important features. Knowledge discovery and data mining, conceptual clustering and the proposed method in this paper are examples of global methods. The method described in evolutionary search uses for evaluation of subsets of features. In EM (Expectation–Maximization) and trace measure are used for evaluation. In hierarchical clustering and conceptual clustering features are ranked and selected for categorical data. Forward and backward search techniques are used to generate candidate subsets. These methods measure the category utility of the clusters by applying COBWEB Knowledge acquisition to evaluate each candidate subset. In document clustering authors proposed an objective function for choosing the feature subset and finding the optimal number of clusters for a document clustering problem using a Bayesian statistical estimation framework. Projected clustering, Automatic subspace clustering, Entropy-based subspace clustering and Projected clustering are examples of local wrapper methods. These methods find subsets of features defining each cluster. ProClus first finds clusters using K-medoid Cluster Analysis considering all features and then finds the most important features for each cluster using Manhattan distance. The algorithm called CLIQUE in Automatic subspace clustering divides each dimension into a user given divisions. It starts with finding dense regions (or clusters) in 1-dimensional data and works upward to find j-dimensional dense regions using candidate generation algorithm. In this paper we proposed a combined filter and wrapper method to evaluate feature subsets and choose the best subset for clustering by considering their effect on the underlying clusters. Earlier methods proposed for clustering were mostly wrapper methods which require some clustering algorithm and some invariant clustering criterion to evaluate feature subsets. A main drawback of this approach is the lack of unanimous agreement in evaluating the clusters. Furthermore, running a clustering

algorithm is very sensitive on some parameters such as the number of clusters or some equivalent of it. For real-world data this information is usually hard to obtain, making it unusable in most cases. But in contrast the proposed method largely depends on a parameter (range of intra-cluster distance) which is easier to set because the proposed method is quite insensitive to it.

II. RELATED WORK

Feature subset selection is the process of identifying and removing as many irrelevant and redundant features as possible because irrelevant features do not contribute to the predictive accuracy, and redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other feature(s).FCBF is a fast filter method which can identify relevant features as well as redundancy among relevant features without pair wise correlation analysis. Of the many feature subset selection algorithms, some can effectively eliminate irrelevant features but fail to handle redundant features, yet some of others can eliminate the irrelevant while taking care of the redundant features. A FAST algorithm, which is using our combined filter and wrapper method, belongs to the second group. Traditionally, feature subset selection research has focused on searching for relevant features. A well known example is Relief, which weighs each feature according to its ability to discriminate instances under different targets based on distance-based criteria function. However, Relief is ineffective at removing redundant features as two predictive but highly correlated features are likely both to be highly weighted. Relief-F extends Relief, enabling this method to work with noisy and incomplete data sets and to deal with multi-class problems, but still cannot identify redundant features. However, along with irrelevant features, redundant features also affect the speed and accuracy of learning algorithms, and thus should be eliminated as well FCBF and CMIM are examples, which consider the redundant features. Different from these algorithms, our proposed combined method employs clustering based method to choose features. Recently, hierarchical clustering has been adopted in word selection in the context of text classification. Distributional clustering has been used to cluster words into groups based either on their participation in particular grammatical relations with other words by Pereira et al. or on the distribution of class labels associated with each word by Baker and McCallum . As distributional clustering of words are agglomerative in nature, and result in sub-optimal word clusters and high computational cost, Dhillon et al. proposed a new information-theoretic divisive algorithm for word clustering and applied it to text classification. Butterworth et al. proposed to cluster features using a special metric of Barthelemy-Montjardet distance, and then makes use of the dendrogram of the resulting cluster hierarchy to choose the most relevant attributes. The cluster evaluation measure based on Barthelemy-Montjardet distance does not identify a feature subset that allows the classifiers to improve their original performance accuracy. Furthermore, even compared with other feature selection methods, the obtained accuracy is lower. Hierarchical clustering also used to select features on spectral data. Van Dijk and Van Hullefor proposed a hybrid filter/wrapper feature subset selection algorithm for regression. Krier et al. presented a methodology combining hierarchical constrained clustering of spectral variables and selection of clusters by mutual information. Their feature clustering method is similar to that of Van Dijk and Van Hullefor except that the former forces every cluster to contain consecutive features only. Both methods employed agglomerative hierarchical clustering to remove redundant features.

III. PRUPOSED APPROACH

The high dimensionality of data can cause data overload, and if there are a lot of features, it is possible that the number of cases in data set to be insufficient for data mining operations. The solution for these problems is the reduction of data dimensions. The size of a data set is determined both by the number of cases and by the number of features considered for each case. In order to reduce number of cases one can use sampling or filtering. Feature reduction may be achieved either by feature composition or by feature selection. These methods should produce fewer features, so the algorithms can learn faster. Sometimes, even the accuracy of built models could be improved. Methods used for feature selection, can be classified as: filters, which are open loop methods, and wrappers, which are closed loop methods.

a) FEATURE SELECTION PROCESS:

The main purpose of Feature selection is to identify and to remove irrelevant and redundant features. It has the potential to be a fully automatic process, and brings some benefits for data mining, such as: an improves predictive accuracy, more compact and easily understood learned knowledge and reduces execution time for algorithms. Feature selection methods are divided in two broad categories, filter method and wrapper method, and within these categories algorithms can be further individualized by the nature of their evaluation function and by the means the space of feature subsets. Feature selection algorithms perform a search through the space of feature subsets, and must solve four problems which affects the search to select a point in the feature subsets space from which to start the search. A first choice, called forward selection, it begin with no features and successively add attributes. The second selection is backward selection, begins with all features and successively remove them, the heuristic search strategies not guaranteed in finding the optimal subset, such strategies can give good results, and are more feasible than exhaustive search strategies which are prohibitive just for a small initial number of features; - the most important factor which differentiate the feature selection algorithms is evaluation strategy. There are feature selections methods which operate independently any learning algorithm, based on general characteristics of the data to evaluate the irrelevant features are filtered before learning begins. An induction algorithm combined with a statistical re-sample technique is used in other methods to estimate the final accuracy of feature subsets.

b) FILTERS VS. WRAPPERS:

The simplest approach to feature selection is filter method, which is called as open loop feature selection methods. Based on class separability criteria, filters do not consider the effect of selected features on the performance of the whole process of knowledge discovery, as is presented in Figure 1(a). They usually provide a ranked list of features that are ordered according to a specific evaluation criterion such as: information content or statistical dependencies between features, accuracy and consistency of data. They also give information about the relevance of a feature compared with the relevance of other features, and do not tell to the analyst what is the desirable minimum set of the features.

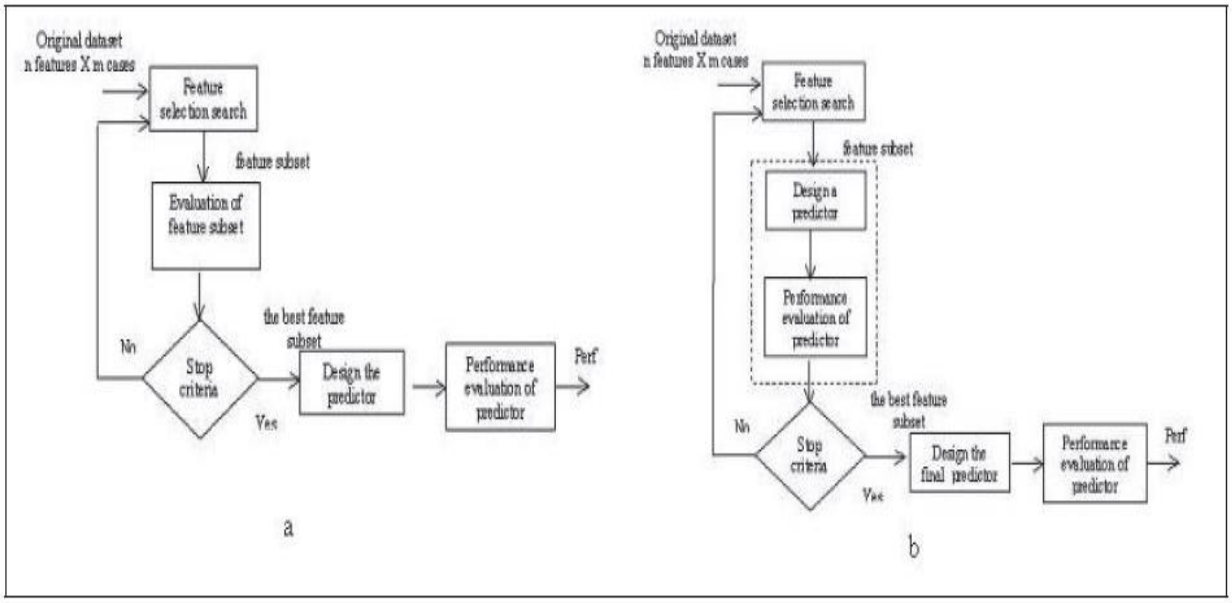

Figure 1: Open loop feature selection method (a), and closed loop feature selection method (b)

Wrappers, known also as closed loop feature selection methods take into account the feedback related to the performance of the selected set of features for the complete KDD process. They use the prediction performance as selection criteria, and evaluate the quality of selected features by comparing the performances for prediction algorithms applied on the reduced set of features and on the original one. Figure 1(b) presents a closed loop feature selection method.

Regarding the final predictive accuracy of a learning algorithm, wrappers often give better results than filters, because feature selection is optimized for the specific learning algorithm used. But if the computational complexity and execution time are considered, wrappers are too expensive for large dimensional datasets, since each selected feature set must be evaluated with the predictive algorithm used.

c) A COMBINED APPROACH FOR FEATURE SELECTION:

Studying the two general feature selection methods we can conclude that, if improved performance for a specific learning algorithm is required, a filter can provide a reduced initial feature subset for a wrapper which contains only relevant features, as shown in Figure 2. This approach could produce shorter and faster search for the wrapper.

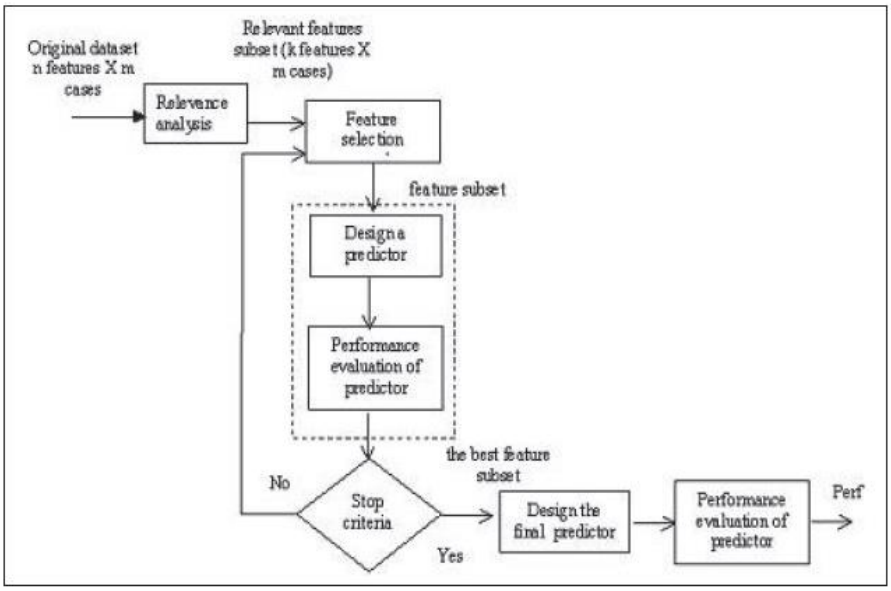

Figure 2: A framework which combines feature relevance analysis with closed loop feature selection

Since practice has demonstrated that irrelevant input features lead to great computational cost for data mining process and may cause over fitting, more feature selection researches focused on extraction of relevant features from the whole data set in order to apply data mining algorithms upon these data. But it is difficult to identify the feature is relevant or not? The

incremental concept formation stated that features are relevant if their values vary systematically with category membership. That means that a feature is relevant if it is correlated with the class. This is defined as follows:
Definition 1. A feature Fi is relevant iff there exists fi and c for which p(Fi = fi) > 0 such that
p(C = c|Fi = fi) 6= p(C = c) (1)
Relevance is usually defined in terms of correlation or mutual information. In order to define mutual information for two features we start from the concept of entropy, as a measure of
uncertainty of a random variable. For a variable X the entropy is defined as:
E(X) = -∑p(xi)log2(p(xi)) (2)
The entropy of a variable X after observing values of another variable Y is defined as:
E(X|Y) = -∑p(yi)∑p(xi, yi)log2(p(xi|yi)) (3)
where p(xi) is the prior probability for all values of X, and p(xi|yi) is the posterior probabilities of X given the value of Y. The value by which the entropy of X decreases, estimates additional information about X provided by Y. It is called information gain and is calculated using the following expression:
I(X, Y) = E(X) − E(X|Y) (4)
We take into account that for discrete random variable, the joint probability mass function is:
p(xi|yj) = p(xi, yj)/p(yj) (5)
and the marginal probability function, p(x) is:
p(xi) = _p(xi, yj) = _p(xi|yj)p(yi) (6)
where p(x,y) is joint probability distribution function of X and Y, and p(xi) and p(yj) are the marginal probability distribution functions of X and Y respectively. Finally, for two discrete random variables X and Y , information gain is formally defined as:
I(X, Y) =∑j∑ip(xi, yj)log(p(xi, yj)/((p(xi)p(yj))) (7)
According to this expression, one says that a feature Y is more correlated to feature X than feature Z if:
I(X, Y) > I(Z, Y) (8)
It can be observed that information gain favors features with more values, so it should be normalized. In order to compensate its bias and to restrict its values to range [0,1] it is preferable to be used symmetrical uncertainty, defined as:
SU(X, Y) = 2 I(X, Y)/(E(X) + E(Y)) (9)
A value of 1 for symmetrical uncertainty means that knowing the values of either feature completely predicts the value of the other whereas a value of 0 implies that X and Y are independent. Starting from these considerations, in the proposed framework, first a relevance analysis is made using the symmetrical uncertainty SU(Fi,C) between each feature Fi and the class C. Based on this analysis one removes the irrelevant features, and one obtains a features subset containing only the relevant features. Then, on this dataset one applies closed loop feature selection methods, using as search strategy both forward and backward selection, and using a decision tree as predictor, both for feature selection and for thse performance evaluation.

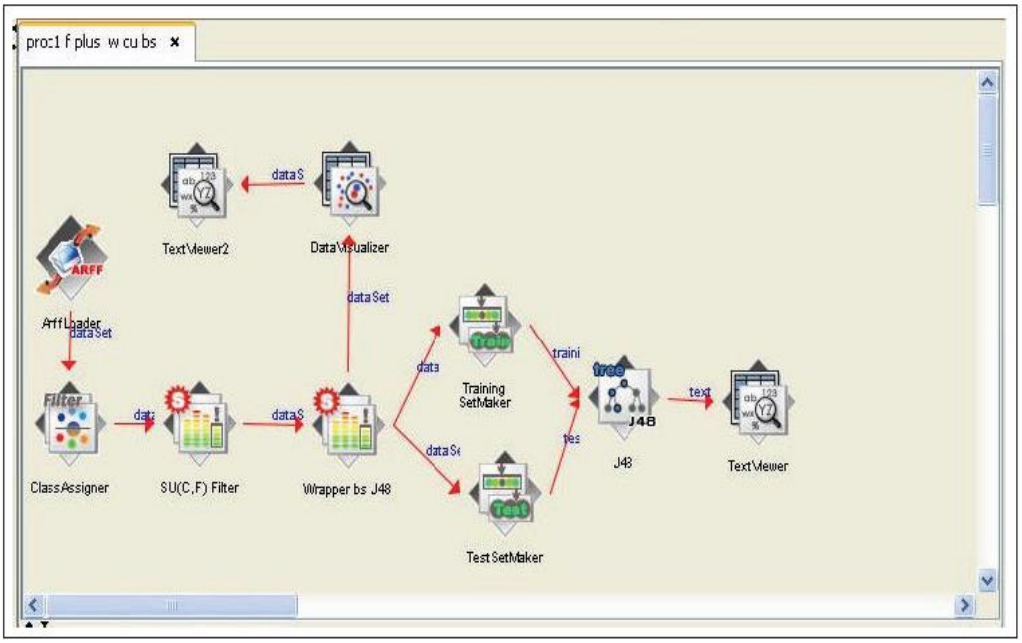

Figure 3: WEKA knowledge flow for the proposed framework

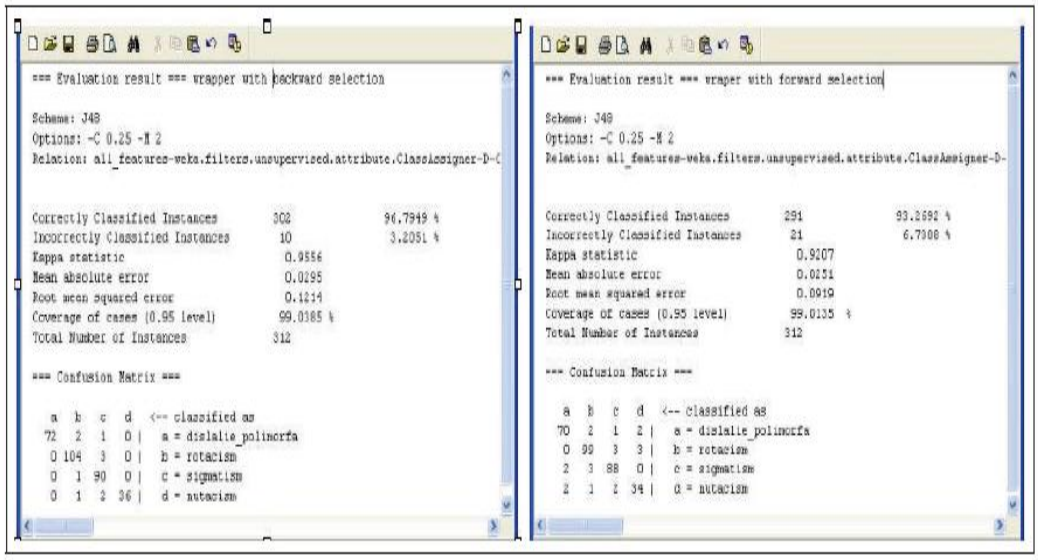

Figure 4: Performances of classifier built on feature subsets obtained by proposed method

IV. CONCLUSION

As a possibility to reduce the number of feature considered by data mining algorithms, in order to make them more efficient, the combined method which uses a combination filter wrapper. The method used a correlation based filter on the whole set of features, then on relevant subset of features it used a wrapper which uses a decision tree classifier for prediction. The combined method achieved clearly superior performances for execution time, when we have used for feature selection the combined approach and backward selection as search strategy for wrapper.

V. ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their constructive comments that helped significantly in improving this paper.

REFERENCES

[1] Liu H. and Setiono R., A Probabilistic Approach to Feature Selection: A Filter Solution, in Proceedings of the 13th International Conference on Machine Learning, pp 319-327, 1996.
[2] Raman B. and Ioerger T.R., Instance-Based Filter for Feature Selection, Journal of Machine Learning Research, 1, pp 1-23, 2002.
[3] Yu L. and Liu H., Feature selection for high-dimensional data: a fast correlation-based filter solution, in Proceedings of 20th International Conference on Machine Leaning, 20(2), pp 856-863, 2003.
[4] Data Dimensionality Reduction for Data Mining: M. Danubianu, S.G. Pentiuc, D.M. Danubianu,INT J COMPUT COMMUN, ISSN 1841-9836 7(5):824-831, December, 2012.
[5] Battiti R., Using mutual information for selecting features in supervised neural net learning, IEEE Transactions on Neural Networks, 5(4), pp 537- 550, 1994.
[6] Almuallim H. and Dietterich T.G., Algorithms for Identifying Relevant Features, In Proceedings of the 9th Canadian Conference on AI, pp 38-45, 1992.
[7] Butterworth R., Piatetsky-Shapiro G. and Simovici D.A., On Feature Selection through Clustering, In Proceedings of the Fifth IEEE international Conference on Data Mining, pp 581-584, 2005.
[8] A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data Qinbao Song, Jingjie Ni and Guangtao Wang, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING VOL:25 NO:1 YEAR 2013
[9] Kohavi R., John G., Wrappers for feature subset selection, Artificial Intelligence, Special issue on relevance, 97(1-2):273-324, 1997.
[10] Yu, L., Liu, H., Efficient Feature Selection via Analysis of Relevance and Redundancy, Journal of Machine Learning Research, 5:1205-1224, 2005 .

AUTHOR PROFIE:

Dr.M.V.Siva Prasad awarded PhD from Nagarjuna University, Guntur, received M.Tech. [SE] from VTU, Belgaum and B.E. [CSE] from Gulbarga University, presently working as principal in Anurag Engineering College (AEC), Ananthagiri(V), Kodad(M), Nalgonda(Dt.), Andhra Pradesh, India.

Ch.Suresh Kumar received Master of Technology (Computer Science & Engineering) from Jawaharlal Nehru Technological University (JNTUH). My research interests include Information Security, Web Services, Cloud Computing, Data Mining and Mobile Computing. Presently working as Associate Professor in CSE Department in Anurag Engineering College (AEC), Ananthagiri (V), Kodad (M), Nalgonda (Dt.), Andhra Pradesh, India.

T.Maneesha Pursuing Master of Technology (Computer Science & Engineering)from Jawaharlal Nehru Technological University (JNTUH). My research interests include Information Security, Web Services, Cloud Computing, Data Mining and Mobile Computing. Presently working as TEQIP Assistant in the department of CSE in Anurag Engineering College (AEC), Ananthagiri (V), Kodad(M), Nalgonda(Dt.), Andhra Pradesh, India.

