FEATURE SELECTION FRAMEWORK BASED ON FILTER MEASURES FOR HIGH DIMENSIONAL DATA

Smita Chormunge[1], Sudarson Jena[2]

[1]Research Scholar, GITAM University, Hyderabad, India
[2]GITAM University, Hyderabad, India

Abstract

The increase of data volume in terms of number of features and instances becomes an immense challenge for feature selection algorithms. It increases the computational cost and decreases the accuracy of learning algorithms. This paper proposes a Feature Selection Comprehensive Framework (FSCF) based on filter measures for high dimensional data to produce optimal feature subset in efficient time. Extensive experiments are carried out to comparison of proposed framework and representative methods with respect to different classifiers like Naive bayes and K-NN classifiers on high dimensional datasets. The results demonstrate that proposed framework not only efficient in computational time but also improve the performance of learning algorithms.

Keywords— *Feature selection; Information gain; filters; naive bayes; k-nearest neighbors; classifiers.*

----------***----------

1. INTRODUCTION

In data mining high dimensional data is big challenging task. Feature selection is one of the solutions to reduce the dimensionality. Feature selection is an active in research and development for decades in statistical pattern recognition and machine learning [8]. Domain experts performed feature selection to determine class label and choose what properties of an object should be measured. In Gene expression tasks, to find the relevant features many thousands of genes are tested and then use a statistical feature selection process for the classification problem. Irrelevant and redundant data is one problem of Feature selection and another is search strategy. Irrelevant features not contain needed information about the classification problem where as redundant features contain duplicate information which is already present in more informative features. In many applications like image retrieval, genomic microarray analysis, text categorization and intrusion detection feature selection has found success. It is also effective in enhancing learning efficiency. There are three general approaches of feature selection: Filters, Wrappers and Embedded methods [6]. Evaluation function is measures of filter and wrapper methods which are classified into three categories distance measures, uncertainty measures and dependence measures. The data intrinsic measures are distance, entropy, and dependence measures. According to recent research evaluation function is divided into five categories: consistency, dependence, distance, information, and classier error rate [7]. Most of the feature selection algorithms [1], [2], [3], [4], [5] have been proposed for classification techniques. Aim of all approaches is to search an optimal feature set and improve the classifier performance. Many of these feature selection algorithms use statistical measures to find relevant and non redundant feature subset it include mutual information, correlation and information gain measure. It also improved search approaches such as population-based heuristic search approach, genetic algorithms and ant colony optimization.

Mitra et.al [11] Proposed unsupervised feature subset selection method to remove redundancy by using feature similarity measure. To calculate similarity between two random variables a new measure called maximal information compression index is used. Krier et al. has proposed a methodology which combined hierarchical constrained clustering of spectral variables and selection of clusters by mutual information measure [12]. Van Dijck et al. presented same methodology as Krier, only difference is that only consecutive features are contain in every clusters [13].

Relief feature selection method introduced by Kira and Rendell it is an instance based attribute ranking scheme [9] and later Relief is enhanced by Kononenko [10]. Relief method sampled instance randomly from the data and then finding its nearest neighbor from the same and opposite class. The sampled instances are compared to the values of the attributes of the nearest neighbors. This is used to update relevance scores for each attribute. M.A.Hall has proposed CFS algorithm to address the problem of feature selection by using correlation measure for machine learning [14]. Song Q et.al [15] proposed a Fast clustering based feature Selection algorithm (FAST) based on the MST method. FAST has a high probability of producing a most relevant subset by using clustering based strategy. Peng [16] introduced a feature selection method called mRMR (MaxRelevance and Min-Redundancy) based on mutual information which minimizes redundancy between features and maximizes dependency among a feature subset and a class label.

This paper proposes the Feature Selection Comprehensive Framework (FSCF) based on Information gain and filter measure for high dimensional data. Combination of these two measures works better to improve efficiency of learning methods. Efficiency concerns computational time to obtain the feature subset. For experimental results we used microarray datasets which features ranges from 200 to more

than 4000. Compared propose framework with other well known feature subset selection methods like Relief and IG with respect to two classifiers Navie bayes and K-nearest neighbor.

The rest of the paper is organized as follows. Section II describes the proposed feature selection comprehensive framework (FSCF) in detail. Section III discussed the Classification methods used for analysis. Empirical study presents in Section IV and experimental results discussed in section V. Finally Section VI of this paper presents the concluding remarks.

2. PROPOSED FEATURE SELECTION COMPREHENSIVE FRAMEWORK (FSCF)

This section discuss proposed framework (FSCF) based on different evaluation measures. While evaluating high dimensional data with well known feature selection algorithm, it takes more time to find feature subset. To overcome this we proposed a feature selection framework based on Information gain and filter measures. This works in two steps process.

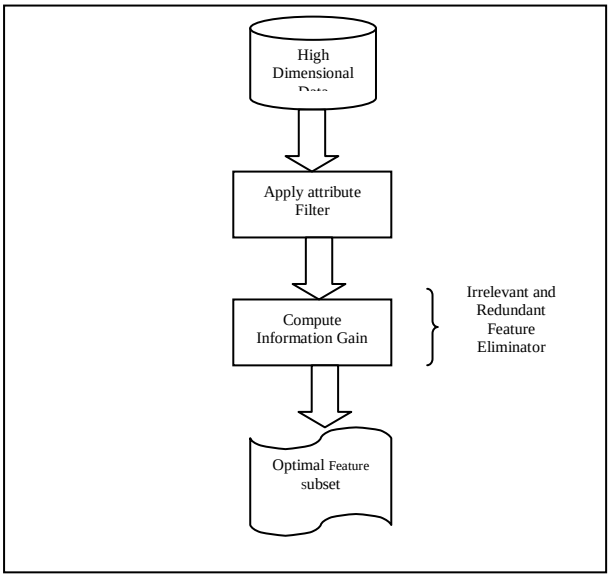

Fig 1. Proposed Feature Selection Comprehensive Framework (FSCF)

First step is applying discrete unsupervised attribute filter on whole datasets F such as F{x_1….x_n}. Second step is to find optimal feature subset using information gain measure.

2.1 Attribute Filter

Discrete unsupervised attribute filter is applied to uniform the data. In machine learning most classification method includes features that are different types such as ordinal, continuous and nominal. Many machine learning algorithms have been developed to deal with mixed data. Discretization is one of the most common methods of accomplishing this task. Discretization filter transform continuous valued attributes to nominal.

Discretization classified into three categories: Supervised versus Unsupervised, Global versus Local and Static versus dynamic. Supervised methods use class label when discretizing features whereas unsupervised discretization approach unknown to class label, for feature it divides the range of observed values into p equal sized bins, where p is user defined parameter. The difference between global and local methods is based on when discretization is performed. Local methods carry out discretization during the induction process where as Global methods discretize features before to induction. Dynamic methods simultaneously look for the space of possible k values for all features. [17].

2.2 Information Gain

Information gain is one of the simplest attribute ranking methods and used in different applications such as text categorization applications where the sheer dimensionality of the data precludes more sophisticated attribute selection techniques [18]. Consider if A is an attribute and C is the class, the entropy of the class before and after observing the attribute is calculated by Equations I and II.

$$H(C) = -\sum_{c\epsilon C} p(c)\log 2p(c) \quad \text{(I)}$$

$$H(C|A) = -\sum_{a\epsilon A} p(a)\sum_{c\epsilon C} p(c|a)\log 2p(c|a)$$

Information gain is calculated by the amount by which the entropy of the class decreases reflects the additional information about the class provided by the attribute [18]. Score of each feature A_i is assigned based on the information gain between itself and the class:

$$IG_i = H(C) - H(C|A_i) \quad \text{(II)}$$
$$= H(A_i) - H(A_i|C)$$
$$= H(A_i) + H(C) - H(A_i, C)$$

3. CLASSIFIERS USED FOR ANALYSIS

3.1 Naive Bayes Classifier

On gene expression data Navie Bayes Classifier provides better classification accuracy than any other classifiers. This classifier learns from training data and then the highest posterior probability predicting the class of the test instance. Consider C ={c_1,c_2,...,c_n} be the set of one classes, and let L ={L_1,...,L_m}be the set of features contained in these classes. If d is a new document, the probability that d belongs to class c_i is defined by Bayes rule,

$$p(d|\,ci) = \frac{p(d|\,ci)p(ci)}{p(d)} \quad \text{(III)}$$

Naive Bayes classifiers are highly scalable. In a learning problem this classifier want a number of parameters linear in the number of features. Evaluating a closed-form expression a Maximum-likelihood training can be completed, it takes linear time than other types of classifiers which used expensive iterative approximation [19].

3.2 K-NN Classifier

The K Nearest-Neighbors is a non-linear classifier. The result of this classifier is depends on whether *k*-NN is used for classification or regression. In this classifier, the output is a class membership. An object is classified based on majority vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors. For example, k is a regularization parameter and x as an example, the k-NN classifier considers the k training examples nearest to x as per their distance in the feature space {0, 1}K and gives as predicted class the dominant real class among those k examples [20].

4. EMPIRICAL STUDY

In this section we discussed the details of datasets. This paper study is on four well known datasets such as Arrhythmia, Colon cancer, SRBCT and Lymphoma [21]. These all are the high dimensional datasets whose features range from 200 to 4000.

Arrhythmia distinguishes between the presence and absence of cardiac arrhythmia and classifies it in one of the 16 groups. This dataset has 452 samples and 279 attributes. Colon Cancer has 62 samples collected from patients. Among 62 samples, 40 samples labeled as negative are tumor biopsies from tumors and 22 samples are positive biopsies are from healthy parts of the colons of the same patients, the total number of genes to be tested is 2000. SRBCT Gene expression data contains 2308 genes for 83 samples from the microarray experiments of Small Round Blue Cell Tumors (SRBCT). Among 83 samples total of 63 are training samples and 25 are test samples. Lymphoma is a broad term encompassing a variety of cancers of the lymphatic system. It contains 4026 number of genes 4026 and 62 number of samples. There are all together three types of lymphomas. The first type, Chronic Lymphocytic Lymphoma, the Follicular Lymphoma is second type and the third type Diffuse Large B-cell Lymphoma [23]. The summary of datasets used for empirical study shown in table I.

Table 1. Summary Of Datasets Used For Empirical Study

Datasets	Features	Instances
Arrithymia	280	452
Colon Cancer	2000	62
SRBCT	2308	83
Lymphoma	4026	62

5. RESULTS AND ANALYSIS

Computational time is calculated for different datasets to obtain feature subset by using FSCF framework. For analyzing results two classifiers used like Naïve bayes and K-NN classifiers, also compared Relief and IG feature selection algorithms with proposed Framework. Weka software is used for evaluating the data. Weka is a data mining tool [22].

Table 2 shows the evaluation of naïve bayes classifiers using FSCF and other two feature selection algorithms IG and Relief. From results we found that FSCF is taking less computational time for evaluating different datasets as compared to other two methods. Whereas IG method performance is better than Relief method for both classifiers. Also calculated the f-measure for all datasets, here almost all these methods shows same accuracy. For Lymphoma dataset time by FSCF is 0.92 sec where as for IG it is 0.88 sec which is less than FSCF but in this case accuracy of FSCF is increases than IG and Relief method.

Table 3 shows the evaluation of K-nearest neighbor classifier on four datasets. In this classifier proposed framework works efficiently than other two methods. Relief method computational time is very large as compared to other two methods, specifically Arrithymia dataset which instances are more in numbers 452 as compare to other datasets. It takes time 8.17 sec for relief method, IG method takes 0.23 sec but FSCF takes very less time as 0.08 sec. Arrithymia dataset have very large instances as 452 because of this there is more variation in time for these datasets. Here FSCF works efficiently even though instances are more in number.

Table 2. Evaluation of Feature selection methods for Naïve bayes classifier

Datasets	FSCF		IG		Relief	
	Time	Fmeasure	Time	Fmeasure	Time	Fmeasure
Arrithymia	0.24	0.623	0.28	0.623	7.91	0.623
Colon Cancer	0.41	0.55	0.58	0.55	1.14	0.55
SRBCT	0.55	0.98	0.64	0.98	4	0.98
Lymphoma	0.92	0.953	0.88	0.936	5.39	0.936

Table 3. Evaluation of Feature selection methods for K-NN classifier

Datasets	FSCF		IG		Relief	
	Time	Fmeasure	Time	Fmeasure	Time	Fmeasure
Arrithymia	0.08	0.505	0.23	0.505	8.17	0.505
Colon Cancer	0.3	0.774	0.17	0.774	1.42	0.774
SRBCT	0.39	0.84	0.47	0.84	2.55	0.84
Lymphoma	0.42	0.97	0.64	0.97	5.38	0.97

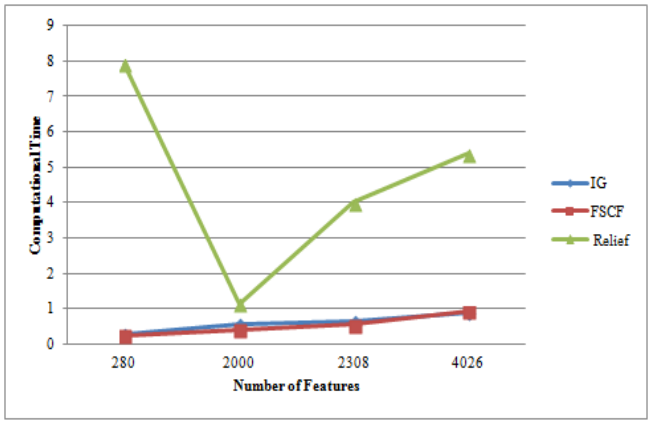

Fig. 2. Comparison graph of feature selection methods for naïve bayes classifiers

Fig.2 and Fig. 3 shows the comparison graph for naïve bayes classifier and comparison graph for K-nearest neighbor classifier respectively. Fig.2 represents the performance of IG, FSCF and Relief method on naïve bayes classifier. This graph represents the computational time over number of features and found the difference of computational time**.** Relief method computational time is too more than other two methods, even same for K-NN classifier. From these results we found that proposed framework FSCF is better in computational time than other two methods for both classifiers. Individual performance of each method on both classifiers is shown in fig. 4, 5 and 6.

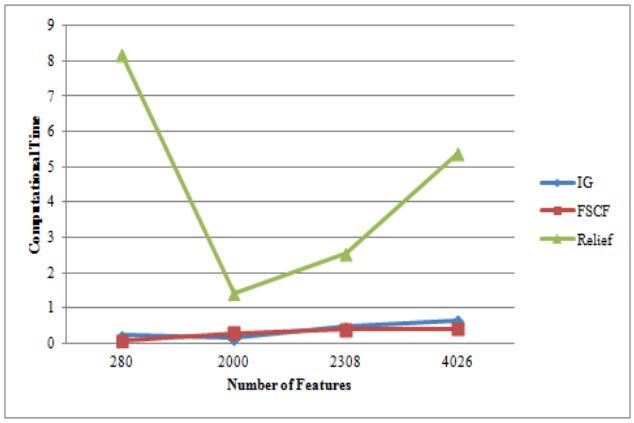

Fig. 3. Comparison graph of feature selection methods for K-NN Classifier

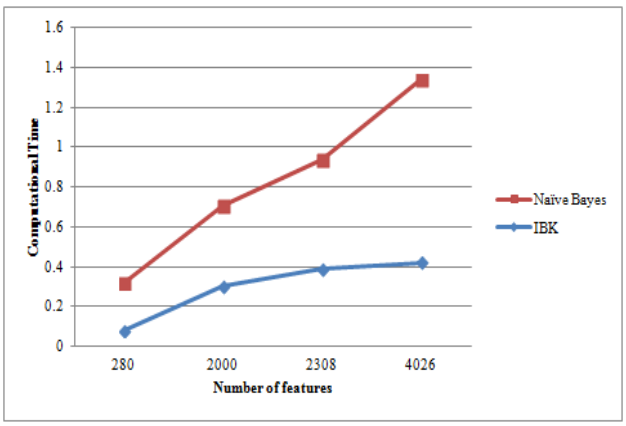

Fig. 4.FSCF performance with Classifiers

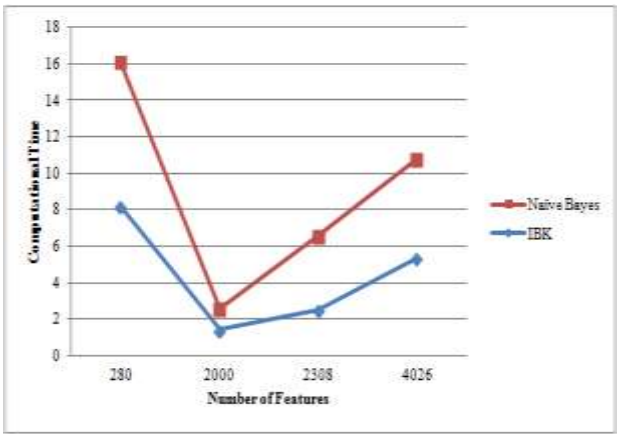

Fig. 5 IG Performance with Classifiers

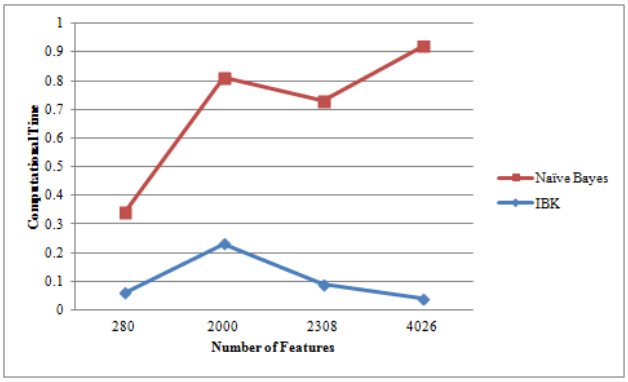

Fig. 6. Relief Performance with Classifiers

FSCF performance with respect to classifiers is shown in figure 4. Fig. 5 and 6 shows the performance of IG and Relief method with respect to classifiers respectively. Classifier accuracy represented in fig. 7. KNN classifier performance is better than Naïve bayes classifier by using proposed framework.

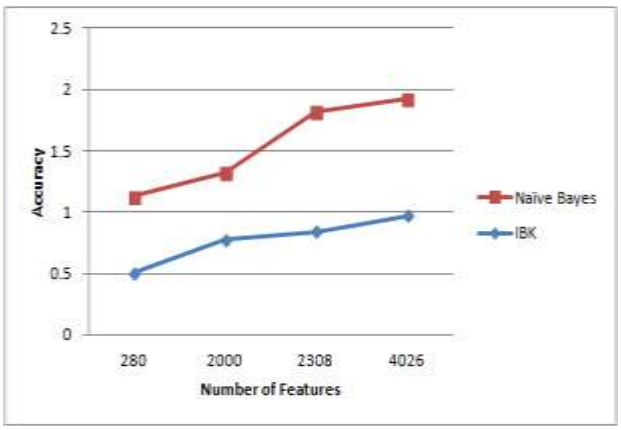

Fig. 7. Classifiers Accuracy with FSCF

6. CONCLUSION

In this paper Feature Selection Comprehensive Framework (FSCF) is proposed based on information gain and filter evaluation measure. Computational time is calculated for finding relevant and non redundant feature subset by using FSCF framework for high dimensional data. Proposed framework compared with other well known feature

selection algorithms like IG and Relief methods on microarray datasets. A result shows that FSCF is efficient in computational runtime than Relief and IG methods with respect to classifiers like naïve bayes and K-NN classifiers. Results may vary based on datasets used for empirical study. Further plan is to improve the accuracy of classifiers by using different measures.

REFERENCES

[1] H. Frohlich, O. Chapelle, B. Scholkopf ," Feature selection for support vector machines by means of genetic algorithm, in: Tools with Artificial Intelligence". Proceedings. 15th IEEE International Conference on, IEEE, pp. 142–148,2003.

[2] S.-W. Lin, K.-C. Ying, C.-Y. Lee, Z.-J. Lee , "An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection", Applied Soft Computing 12 (10) 3285–3290,2012.

[3] L. Yu, H. Liu , "Redundancy based feature selection for microarray data", Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2004, pp. 737–742,2004.

[4] D. K. Bhattacharyya, J. K. Kalita, "Network Anomaly Detection: A Machine Learning Perspective", CRC Press,2013.

[5] S. Nemati, M. E. Basiri, N. Ghasem-Aghaee, M. H. Aghdam , "A novel aco–ga hybrid algorithm for feature selection in protein function prediction", Expert systems with applications 36 (10) 12086–12094,2009.

[6] I. Guyon and A. Elisseeff , "An Introduction to Variable and Feature Selection," J. Machine Learning Research, vol 3, pp. 11571182,2003.

[7] M. Dash and H. Liu , "Feature Selection for Classification," Intelligent Data Analysis, vol. 1, no. 3, pp. 131-156,1997.

[8] L. Yu and H. Liu , "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," Proc. 20th Int'l Conf. Machine Leaning, vol. 20, no. 2, pp. 856-863, 2003.

[9] K. Kira and L.A. Rendell, "The Feature Selection Problem:Traditional Methods and a New Algorithm," Proc. 10th Nat'l Conf.Artificial Intelligence, pp. 129-134, 1992.

[10] I. Kononenko , "Estimating Attributes: Analysis and Extensions of RELIEF," Proc. European Conf. Machine Learning, pp. 171-182,1994.

[11] P. Mitra, C. Murthy, S. K. Pal, "Unsupervised feature selection using feature similarity", IEEE transactions on pattern analysis and machine intelligence 24 (3) 301–312,2002.

[12] C. Krier, D. Francois, F. Rossi, and M. Verleysen, "Feature Clustering and Mutual Information for the Selection of Variables in Spectral Data," Proc. European Symp. Artificial Neural Networks Advances in Computational Intelligence and Learning, pp. 157-162, 2007.

[13] G. Van Dijck and M.M. Van Hulle , "Speeding Up the Wrapper Feature Subset Selection in Regression by Mutual Information Relevance and Redundancy Analysis," Proc. Int'l Conf. Artificial Neural Networks,2006

[14] M.A. Hall , "Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning," Proc. 17th Int'l Conf. Machine Learning, pp. 359-366,2000.

[15] Qinbao Song, Jingjie Ni and Guanta Wang ," Fast Clustering Based Feature Subset Selection Algorithm for High Dimensional Data", IEEE Transactions on Knowledge and Data Engineering,Vol 25, No.1,pp-1-14,2013.

[16] Hanchuan Peng, Fuhui Long,Chris Ding, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy", IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL. 27, NO. 8,pp 1226-1238,2005.

[17] M.A. Hall , "Correlation-Based Feature Subset Selection for Machine Learning," PhD dissertation, Univ. of Waikato,1999.

[18] Mark A. Hall, Geoffrey Holmes, "Benchmarking Attribute Selection Techniques for Discrete Class Data Mining", IEEE Transactions on Knowledge and Data Engineering, VOL. 15, NO. 3,2003.

[19] I.S. Dhillon, S. Mallela, and R. Kumar, "A Divisive Information Theoretic Feature Clustering Algorithm for Text Classification," J. Machine Learning Research, vol. 3, pp. 1265-1287,2003.

[20] Mark A. Hall, Geoffrey Holmes,F. Fleuret , "Fast Binary Feature Selection with Conditional Mutual Information," J. Machine Learning Research, vol. 5, pp. 1531-1555,2004.

[21] Datasets can be downloaded from: http://repository.seasr.org/Datasets/UCI/arff/arrhythmia.arff, http://repository.seasr.org/Datasets/UCI/arff/.

[22] Remco R. Bouckaert,Eibe Frank,Mark Hall,Richard Kirkby,Peter Reutemann,Alex Seewald,David Scuse , WEKA Manual for Version 3-7-10,July 31 2013.

[23] Chormunge Smita and Sudarson Jena, "Metric Based Performance Analysis of Clustering Algorithms for High Dimensional Data", IEEE Fifth International Conference on Communication Systems and Network Technologies, pp-1060-1064,2015.