IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.7, Special Issue.3, Feb. 2019 E-ISSN: 2347-2693

Feature Selection on High Dimensional Big Data of Gens Expression Using Filter Based Feature Selection Methods

A. K. Shrivas[1*], Prem Kumar Chandrakar[2]

[1]Department of Computer Science, Mahant Laxmi Narayan Das College, Raipur, India
[2]Department of Information and Technology, Dr. C. V. Raman University, Bilaspur, India

[*]*Corresponding Author: akhilesh.mca29@gmail.com*

Available online at: www.ijcseonline.org

***Abstract*—** Feature selection approach solves the dimensionality problem by removing irrelevant and redundant features. Recently, big data is widely available in information systems and data mining has pulled in a major thoughtfulness regarding analysts to transform such information into helpful learning. This implies the presence of low quality, questionable, excess and uproarious information which contrarily influence the way toward watching learning and helpful example. As follows, researchers require related big data utilizing feature selection methods. The process of feature selection is identifying the most relevant attributes and removing the redundant and irrelevant attributes. In this paper, find out the result of different feature selection methods based on a recognized dataset (i.e., gens expression dataset) and classification algorithms were used to evaluate the performance of the algorithms. In this study revealed that feature selection methods are capable to improve the performance of learning algorithms. Still, there are no any single filter based feature selection method is the best. Taken as a whole, Classifier AttEval, Correlation AttributeEval, Principal Components, and ReliefAttEval methods performed better results than the others.

***Keyword*—**Feature selection, Lung cancer, Gens expression, Classifier, Subset.

I. INTRODUCTION

Gene expression data set producing huge amounts of data. This measure of data imply low quality, unreliable, redundant and noisy data to examine useful pattern (Ashraf, Chetty, & Tran, 2013). Therefore, researchers require relevant and high quality data from big data using feature selection methods. Feature selection methods reduce the dimensionality of feature space, remove redundant, irrelevant or noisy data. It brings the immediate effects for application: speeding up a data mining algorithm, improving the data quality and the performance of data mining and increasing the comprehensibility of the mining results(Novaković, Strbac, & Bulatović, 2011) . In this study, the lung cancer disease was considered which is a serious health problem in the world and a comparative analysis of several filter based selection algorithms was carried out based on the performance of classification algorithms for the prediction of disease risks(Yasin, 2011). The main aim of this study is to make contributions in the prediction of lung cancer disease for medical research and introduce a detailed and comprehensive comparison of popular filter based feature selection methods.

II. FEATURE SELECTION METHODS

Several feature selection methods have been introduced in the machine learning domain. The main aim of these techniques is to remove irrelevant or redundant features from the dataset. Feature selection methods have two categories: wrapper and filter. The wrapper evaluates and selects attributes based on accuracy estimates by the target learning algorithm. Using a certain learning algorithm, wrapper basically searches the feature space by omitting some features and testing the impact of feature omission on the prediction metrics. The feature that make significant difference in learning process implies it does matter and should be considered as a high quality feature. On the other hand, filter uses the general characteristics of data itself and work separately from the learning algorithm. Specifically, filter uses the numerical relationship among a set of features and the target feature. The amount of correlation between features and the target variable determine the importance of target variable(Ashraf et al., 2013), (Leach, 2012). Filter based approaches are not dependent on classifiers and usually faster and more scalable than wrapper based methods. In addition, they have low computational complexity.

A. Relief

With the help of Relief-F feature selection method evaluates a feature basically instance based feature which is selected a feature by how well its value distinguish samples that are from dissimilar groups but are similar to each other (Lee, Lushington, & Visvanathan, 2011).

B. B. One-R

One-R is a simple algorithm proposed by (Holte, 1993). One-R algorithm create one rule for each attribute in the training data and then selects the rule with the smallest error. One-R classification produce statistical valued features as continuous and uses a straight forward method to divide the range of values into several disjoint intervals (Novaković et al., 2011).

C. Principal Component Analysis (PCA)

Large number of correlated attributes reduces the dimensionality of dataset using PCA. PCA contains a by transforming the original attributes space to a new space in which attributes are uncorrelated. The algorithm then ranks the variation between the original dataset and the new one (Ashraf et al., 2013), (Jolliffe, 2002).

D. Correlation Based Feature Selection (CFS)

A ranks feature subsets and discovers the merit of feature or subset of features according to a correlation based heuristic evaluation function using CFS filter algorithm.CFS provide a ranking based features to find out subsets that contain features that are highly correlated with the class and uncorrelated with each other (Hall, 1999).

E. Consistency Based Subset Evaluation (CS):-

The class consistency rate is evaluation by CS. CS obtain a set of attributes that divide the original dataset into subsets that contain one class majority(Hall, 1999). One of well known consistency based feature selection is consistency metric proposed by (Liu, Setiono, Science, & Ridge, 1995).

III. CLASSIFICATION ALGORITHMS

An extensive variety of classification algorithms is accessible, each with its qualities and weaknesses. There is no single learning algorithm that works best on all supervised learning issues. This area gives a short outline of four directed learning calculations utilized as a part of this investigation, specifically, J48, Naïve Bayes, IBK and Decision.

A. J48

J48 is the Weka implementation of the C4.5 algorithm, based on the ID3 algorithm. The primary thought is to make the tree by using the information entropy. For every node the most effectively split criteria is calculated and then subsets are generated. To get the split criteria the algorithm looks for the attribute with highest normalized information gain.

B. Naïve Bayes

One of most important algorithm that are based on probability is a naive bayes algorithm that calculates a set of probabilities by together with the frequency and combinations of values in a given data set. Naive Bayes algorithm use Bayes theorem and assume all attributes to be independent given the value of the class variable (Patil, 2013),(Dimitoglou, Adams, & Jim, 2012).

C. IBK

IBK is a case based learning approach like the K-closest neighbor technique. The essential rule of this calculation is that every unseen example is constantly contrasted and existing ones utilizing a separation metric most regularly Euclidean separation and the nearest existing case are utilized to relegate the class for the test (Witten, Frank, & Hall, 2011).

D. Decision Table

Using decision Table data set is summarize with a decision table, the same number of attributes of original dataset is under the decision table, and decision table find out the a new data which is matches the non-class values of the data item.(Kohavi & John, 1997), (A. Tsymbal and S. Puuronen, 2010).

IV. LITERATURE REVIEW

There are a few examinations in view of data mining of biomedical datasets in the literature. Sathyadevi et al., used CART, C4.5 and ID3 algorithms to diagnose lung cancer disease effectively. Agreeing their outcomes, CART calculation performed best outcomes to recognize to disease (Sathyadevi, 2011). Roslina et al. used Support Vector Machines to foresee lung disease and utilized wrapper based component determination technique to recognize important highlights previously classification. . Combining wrapper based methods and Support vector machines produced good classification results (Roslina & Noraziah, 2010).Huang et al. connected a channel based component determination strategy utilizing irregularity rate measure and discretization, to a restorative cases database to anticipate the sufficiency of span of energizer medication use. They utilized strategic relapse and choice tree calculations. Their outcomes recommend it might be practical and effective to apply the channel based element choice strategy to decrease the dimensionality of healthcare databases (Huang, Wulsin, Li, & Guo, 2009). Inza et al. researched the urgent errand of precise quality choice in class forecast issues over DNA microarray datasets. They utilized two surely understood datasets associated with the determination of cancer such as Colon and Leukemia. The outcomes featured that channel and wrapper based quality determination approaches prompt extensively better precision brings about correlation with the non-gene selection system, combined with intriguing and

striking dimensionality reductions (Inza, Larrañaga, Blanco, & Cerrolaza, 2004).

V. DATASETS

Gens expression lung cancer data set contain total of 203 snap-frozen lung tumors (n _ 186) and normal lung (n _ 17) specimens were used to create two datasets. Complete data set contain, 125 adenocarcinoma samples were associated with clinical data and with histological slides from adjacent sections. Lung cancer data set of 203 specimens (Dataset A) include histologically defined lung adenocarcinomas (n _ 127), squamous cell lung carcinomas (n _ 21), pulmonary carcinoids (n _ 20), SCLC (n _ 6) cases, and normal lung (n _ 17) specimens. Another adenocarcinomas (n _ 12) were suspected to be extrapulmonary metastases based on clinical history (see SampleData.xls, which is published as supporting information on the PNAS web site, www.pnas.org, and at www. genome.wi.mit.edu_MPR_lung). Dataset B, a subset of Dataset A, includes only adenocarcinomas and normal lung samples (Bhattacharjee et al., 2001).

The following encoding of diagnostic categories is used:

Table 1. Lung Cancer Dataset

Adeno	0
Normal	1
Squamous	2
COID	3
SMCL	4

VI. EXPERIMENTAL RESULTS

Lung cancer dataset was used to compare different filter based feature selection methods for the prediction of disease risks. Four classification algorithms reviewed above were considered to evaluate classification accuracy. The feature selection methods are Cfs Subset Eval (GeneticSearch),Cfs SubsetEval(GreedyStepwise),CfsSubsetEval(BestFirst),Clas sifierAttEval(ZeroR),CorrelationAttributeEval(Ranker),Prin cipalComponents(Ranker),ReliefAttEval(Ranker).At first, feature selection methods were used to find relevant features in the lung cancer dataset and then, classification algorithms were applied to the selected features to evaluate the algorithms. Respectively 5667, 74, 78 ,75,54, 54 and 74 features were selected by the feature selection algorithms. Same experiment was repeated for four classifiers. WEKA 3.6.8 software was used. WEKA is a gathering of machine learning calculations for data mining tasks and is open source software. WEKA software contain tools for data pre-processing, feature selection, classification, clustering, association rules and visualization (Jolliffe,2002),(http://www.cs.waikato.ac.nz/ml/weka, n.d.).Table shows the performance of features which was selected by the feature selection methods. According to table, the highest precision values were obtained for the lung cancer dataset with Decision Table classifiers with Attribute Eval, One-R Attribute Eval and Relief Attribute Eval.

Table 2. Evaluation of Feature Selection Methods for Lung Cancer Dataset

No. of features in original gens expression dataset	**Attribute Evaluator**	**Search Method /Classifier**	**No. of selected features**
12600	Cfs Subset Eval	GeneticSearch	5667
12600	Cfs Subset Eval	GreedyStepwise	74
12600	Cfs Subset Eval	BestFirst	78
12600	Classifier AttEval	ZeroR	75
12600	Correlation AttributeEval	Ranker	54
12600	Principal Components	Ranker	54
12600	ReliefAttEval	Ranker	74

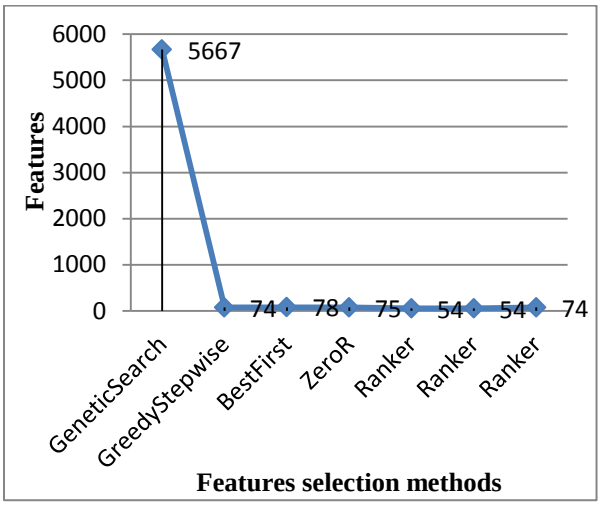

Figure 1. Number of selected features

VII. CONCLUSION

The genes expression lung cancer data set contain 12600 fields of 203 sample data so feature selection is an important data processing step in data mining studies and many machine learning algorithms can hardly cope with large amounts of irrelevant features. Thus, feature selection approaches became a necessity for many studies. In this

study, a comparative analysis was carried out on the basis of filter based feature selection algorithms to predict the risks of lung cancer disease. Six feature selection algorithms were used to analyze the dataset and their performance was evaluated by using J48, Naïve Bayes, IBK and Decision Table classifiers. Among the algorithms, Naïve Bayes and Decision Table classifiers have higher accuracy rates on the lung cancer dataset than the others after the application of feature selection methods. In this study asserted that feature selection methods are capable to improve the performance of learning algorithms. However, no single filter based feature selection method is the best. Overall, Consistency SubsetEval, InfoGain AttributeEval, OneRAttributeEval and ReliefAttributeEval methods performed better results than the others. The results of this study can make contributions in the prediction of lung cancer disease in medical research and provide a deep comparison of popular filter based feature selection methods for machine learning studies. As a future work, a study will be planned to investigate the effects of both continuous and discrete attributes of medical datasets in the performance of feature selection methods and classification accuracy.

REFERENCES

[1] A. Tsymbal and S. Puuronen. (2010). Local feature selection with dynamic integration of classifiers. *Foundations of Intelligent Systems*, 363–375.

[2] Ashraf, M., Chetty, G., & Tran, D. (2013). Feature Selection Techniques on Thyroid , Hepatitis , and Breast Cancer Datasets, *3*(March), 1–8.

[3] Bhattacharjee, a, Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., … Meyerson, M. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A*, *98*(24), 13790–5.

[4] Dimitoglou, G., Adams, J. a, & Jim, C. M. (2012). Comparison of the C4.5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Survivability. *Journal of Neural Computing*, *4*(8), 1–9.

[5] Hall, M. (1999). Correlation-based Feature Selection for Machine Learning. *Methodology*, *21i195-i20*(April), 1–5.

[6] Holte, R. C. (1993). Very Simple Classi fi cation Rules Perform Well on Most Commonly Used Datasets. *Machine Learning*, *11*(1), 63–91.

[7] http://www.cs.waikato.ac.nz/ml/weka. (n.d.). WEKA: Weka 3: Data Mining Software in Java.

[8] Huang, S. H., Wulsin, L. R., Li, H., & Guo, J. (2009). Dimensionality reduction for knowledge discovery in medical claims database: Application to antidepressant medication utilization study. *Computer Methods and Programs in Biomedicine*, *93*(2), 115–123

[9] Inza, I., Larrañaga, P., Blanco, R., & Cerrolaza, A. J. (2004). Filter versus wrapper gene selection approaches in DNA microarray domains. *Artificial Intelligence in Medicine*, *31*(2), 91–103.

[10] Jolliffe, I. T. (2002). Principal Component Analysis, Second Edition. *Encyclopedia of Statistics in Behavioral Science*, *30*(3), 487.

[11] Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1–2), 273–324.

[12] Leach, M. (2012). Parallelising feature selection algorithms. *University of Manchester*.

[13] Lee, I.-H., Lushington, G. H., & Visvanathan, M. (2011). A filter-based feature selection approach for identifying potential biomarkers for lung cancer. *Journal of Clinical Bioinformatics*, *1*(1), 11.

[14] Liu, H., Setiono, R., Science, C., & Ridge, K. (1995). Chi2: Feature Selection, 388–391.

[15] Novaković, J., Strbac, P., & Bulatović, D. (2011). Toward optimal feature selection using ranking methods and classification algorithms. *Yugoslav Journal of Operations Research*, *21*(1), 119–135.

[16] Patil, T. R. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. *International Journal Of Computer Science And Applications, ISSN: 0974-1011*, *6*(2), 256–261.

[17] Roslina, A. H., & Noraziah, A. (2010). Prediction of hepatitis prognosis using support vector machines and wrapper method. *Proceedings - 2010 7th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2010*, *5*(Fskd), 2209–2211.

[18] Sathyadevi, G. (2011). Application of CART algorithm in hepatitis disease diagnosis. *International Conference on Recent Trends in Information Technology, ICRTIT 2011*, 1283–1287.

[19] Witten, I. H., Frank, E., & Hall, M. a. (2011). *Data Mining: Practical Machine Learning Tools and Techniques (Google eBook). Complementary literature None*.

[20] Yasin, H. (2011). Hepatitis-C Classification using Data Mining Techniques, *24*(3), 1–6.

Authors Profile

Mr. Prem Kumar Chandrakar is working as Assistant Professor in Department of Computer Science, Mahant Laxmi Narayan Das College, Raipur, Chhattisgarh, India. He obtained his Master's Degree in Computer Application from Chhattisgarh Swami Vivekanand Technical University, Durg, India and M. Phil in Computer Science from Pt. Ravishankar Shukla University, Raipur, India. He has more than 8 years teaching and 03 years research experience. He has published more than 5 research papers in reputed journals and attended workshop and conference at national and international level. His area of interest includes data mining and big data.

Dr. Akhilesh Kumar Shrivas is working as Assistant Professor in Department of Information Technology, Dr. C.V. Raman University, Bilaspur, India. He obtained his Master's Degree in Computer Application from Guru Ghasidas Vishwavidyalaya, Bilaspur, India and Ph. D. in Computer Science from Dr. C.V. Raman University, Bilaspur, India. He has 6 year research experience and published more than 50 research papers in reputed journals and conference proceedings and attended workshop and conference at national and international level. His research interests are data mining, soft computing, big data and information security.

