ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
DOI: 10.17485/ijst/2015/v8i8/64508

Indian Journal of Science and Technology, *Vol 8(8), 695–701, April 2015*

Gene Selection Using Information Theory and Statistical Approach

Kaberi Das[1], Jagannath Ray[1] and Debahuti Mishra[2*]

[1]Computer Applications, Institute of Technical Education and Research, Siksha O Anusandhan University, Bhubaneswar, Odisha, India.
[2]Computer Science & Engineering, Institute of Technical Education and Research, Siksha O Anusandhan University, Bhubaneswar, Odisha, India; debahutimishra@soauniversity.ac.in

Abstract

This paper focuses on a methodological framework for gene selection by two approaches such as statistical approach and information based approach. Statistical measures are univariate measures where the gene relevance score of each gene is calculated without considering its co-relation (positive co-relation or negative co-relation) with other genes. Statistical approach includes Euclidian distance and Pearson co-relation. Mutual information is the measure of mutual dependence between two random variables in the case of probability theory. Information based approach includes information gain and dynamic relevance. In this paper the above gene selection methods are applied on four publicly available data sets such as, breast cancer, leukemia, hepatitis and dermatology to generate the subset of genes. Then, the resultant subset is fed through two classifiers namely Naive-Bayes and Support Vector Machine (SVM). Here also the data sets are directly applied to the classifier without applying the gene selection methods. Finally when we have compared the result, it has been found that all the data sets showing better accuracy when the classifiers are applied after gene selection technique which reflects the importance of gene selection technique.

Keywords: Information Based Approach, Naïve Bayes, Statistical Approach, Support Vector Machine (SVM)

1. Introduction

Disease diagnosis is a step by step procedure which includes studying medical history, physical examination, clinical tests and so on. To find the disease or diseased part is a long term process due to which the patient has to suffer for a long period. This problem can be addressed by fast diagnosis of diseases and subsequent medications. To achieve fast diagnosis we can apply gene selection approach upon microarray dataset. Studying and analyzing a number of datasets for a particular disease is next to impossible for a human, hence with the fastness characteristics of computers we can get results in minimum amount of time. This is because of the fact that all the genes are not that much informative to recognize the state of a disease[1]. There are only few genes which provide relevant information. This problem is solved by *gene selection* where certain conditions and constraints can be written in an algorithmic way to select a few but most informative genes that can be used on some data-mining tools like classifiers to correctly classify the dataset. The main drawback is getting 100% accuracy on the classification is most of the time not possible and it depends up on the selection technique[2,3]. Hence to get better accuracy we have to supply the best set of informative genes to the classifier.

There are two classical approaches to gene selection namely wrapper and filter[4]. Wrapper approach is classifier specific. Broadly it can be viewed as an approach where the algorithm iteratively selects a set of genes which can be applied on a particular classifier. This process continues till we obtain a selected subset which gives the best accuracy for that particular classifier[5]. Hence, the name wrapper, as here the classifier is wrapped

*Author for correspondence

around the selection algorithm. Filter method on the other hand uses the criterions for selection of particular subset which then can be applied on a number of classifiers to find the classifier which gives the best result. Filter method is mostly preferred over wrapper approach due to its lower computational aspect and time complexity. Although wrapper method has good classification accuracy, sometimes filter methods also gives comparable accuracies.

Analyzing micro array datasets in different works have produced good classification accuracy and to determine which is the disease and which can be the options for treatment. Some of the works with respect to gene selection strategies has been discussed here. A new filter based selection measure has been proposed by Xin Sun *et al.*[4], which is based on information theory. The proposed method Dynamic Relevance (DR) Analysis is an extensive extension of the mRMR approach where the relevance of each gene is updated with iteration where an interdependent gene is selected into the selected subset. The relevance value of the remaining genes is updated and recalculated as the relevance between the newly selected gene and the remaining gene. C.O. Sakar *et al.*[6] introduced a good overview of the minimum Redundancy Maximum Relevance (mRMR) approach. This strategy focuses on the fact that individually good genes do not necessarily gives good classification accuracy like a group. Thus to improve the joint classification accuracy the redundancy among them should be reduced. This approach is based on the information theory. They have proposed an improved approach on mRMR called as Kernel Canonical Correlation Analysis mRMR (KCCAmRMR) where during the calculation of mutual information, instead of the feature X_i the correlated function $fiu(X_i)$ is used. Where $fiu(X_i)$ denotes the various relations of X_i with target class T. Thus we perform a filtering out process to remove all the irrelevant relations. Monalisa Mandal and Anirban Mukhopadhya[7] have proposed an improvement to the existing mRMR approach. In their experimental evaluation on feature selection, the authors have selected the features based on maximizing the relevance between the feature and the class and then minimizing the redundancies between the feature and the other features. Then the feature which is most relevant is selected to the output set and another solution set is created comprising of the rest of the features based on the two selection criteria. Subsequently some features which satisfy both criteria are selected into the final set. The number of features in the final set is to be provided by the user. Yibing Chen *et al.*[8] introduced the feature selection in two phases. In the first phase, they used Bhattacharyya distance to separate the non-informative genes to create a smaller set of informative genes. In the second phase authors used kernel distance as a strategy to measure the class separability which is a way of Floating Sequential Search Method (FSSM). They have applied this on a colon cancer dataset with SVM as the classifier and achieved worthy results. Subhra Sankar Ray *et al.*[9] proposed a new distance measure named as 'Maxrange Distance' to compute similarity between two genes. In this approach normalization is first done on the distance between two genes. The normalizing factor is different for the different experiments which give the data-set although it is similar for all the genes in the data-set. They took the normalizing factor as the 'linear dynamic range' of the 'Photo Multiplier Tube' which is used to scan the 'Fluorescence intensities' of that experiment. Selwyn Piramuthu[10] used the 'Hausdorff Distance' for feature selection which is the measure of similarity between two features in metric space. The distance is calculated between the features of two different classes. Then a decision tree is constructed by taking the distances in ascending order. The tree stops on a stopping criterion and then the features are sent to the selected subset. The quality of the tree is evaluated by classifying unseen examples. Hui-Huang Hsu and Ming-Da Lu[11] in their study used two different approaches to feature selection approaches i.e. Euclidian distance and Pearson Correlation Coefficient both of which are statistical measures.

The outline of this paper is as follows: Section 2 deals with different methods for gene selection, in section 3 schematic representations of proposed model, section 4 deals with the implementation with result analysis and finally section 5 includes the conclusion and future works.

2. Gene Selection Methods

2.1 Information Theory

Information theory is the branch of mathematics and computer science which is related to quantifying the information of a random variable and also to find the information shared between two random variables. For these purposes it uses the concept of Entropy and Mutual Information[4]. Entropy is the measure of uncertainty of

a random variable. The term refers to Shannon entropy. Entropy *H(X)* of a random variable *X* can be defined as:

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \quad (1)$$

Where *p(x)* is the probability distribution function of the random variable *X* and can be defined as

$$p(x) = p(X == x) \quad (2)$$

NB. Here *X* is a discrete random variable. The joint entropy of two random variables '*X* and *Y*' and thus the conditional entropy can be defined as in equation 3, 4 respectively:

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y) \quad (3)$$

$$H\left(X/Y\right) = -\sum_{x \in X} \sum_{y \in Y} p\left(x/y\right) \log p\left(x/y\right) \quad (4)$$

Mutual information of two random variables is the measure of dependence between two random variables. As it is the mutual information between two discrete random variables which is defined as:

$$I(X, Y) = -\sum_{x \in X} \sum_{y \in Y} \frac{p(x, y) \log(p(x, y))}{p(x), p(y)} \quad (5)$$

Mutual information can also be expressed as conditioned on a third discrete random variable 'Z'. It can be expressed in terms of entropy as:

$$I\left(X; Y/Z\right) = H\left(X/Z\right) - H\left(X/Y, Z\right) \quad (6)$$

2.2 Statistical Measures

Statistical measures are univariate measures where the gene relevance score of each gene is calculated without considering its co-relation (positive co-relation or negative co-relation) with other genes[12]. It includes two popular methods.

2.2.1 Euclidean Distance

Euclidean distance is the distance between two points and is given in terms of Pythagorean formula. It is the square root of square of difference between corresponding co-ordinates of individual genes in metric space[13,14]. If there are two genes given by:

$$p = (p_1, p_2, \ldots, p_n) \text{ and } q = (q_1, q_2, \ldots, q_n) \quad (7)$$

Then Euclidian distance between P and Q is:

$$d(P, Q) = d(Q, P) = \sqrt{(Q_1 - P_1)^2 + (Q_2 - P_2)^2 + \ldots + (Q_n - P_n)^2} = \sqrt{\sum_{i=1}^{n} (Q_i - P_i)^2} \quad (8)$$

It is a simple statistical approach based on distance or closeness measure between two or more genes. In the method of Euclidian distance the distance between the genes is calculated. For this we first need to separate the genes according to their given class value since the distance is calculated between the genes of the same class. This is stored in a set which contains the distance between corresponding genes. The total number of distances calculated is an arithmetic sequence or progression with difference of 1. It is given by the formula (9).

$$= (n-1) + (n-2) + (n-3) + \cdots + (n-(n-1)) \quad (9)$$

Then a user specified number of least distances are stored in a separate set (mat) and the genes between which these distance are found are stored in another set (newmat) in the order as they appear in the order in the least distance set (mat). From this set (newmat), a user specified number of genes is selected into our selected subset. A concept of 'PER' is used here. PER is the percentage of the number of candidates to be taken into consideration during selection into the sets 'mat' and 'newmat'. A ceiling function is used when the PER value comes in decimal. From the initial set of distance, PER is the limit of the number of minimum distances to be selected into another set. The respective indices of the distances are stored in the set 'mat'. On this set, a number of iterations, which is a multiple of the PER value, is run to select the final gene set.

2.2.2 Pearson Co-efficient

Pearson product-moment co-relation coefficient is a measure between two variables X and Y, which mathematically calculates by how much they are co-related. The calculated value falls in the range +1 to –1 both inclusive, where +1 denotes total positive correlation or complete co-relation, 0 represents no correlation, and –1 is total negative correlation. In area of data-mining, positive co-relation denotes that presence of one variable substantially increases the classifying power of the other while negative co-relation does just the opposite. This measure was introduced by Karl Pearson[13]. This measure can be applied on a gene expression data-set provided that all the

attributes have numeric values. The Pearson co-efficient is given by (10).

$$\gamma_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - (N\overline{AB})}{N\sigma_A\sigma_B} \quad (10)$$

Where, '*N*' represents the number of attributes counted from i = 1: n, '*A*' and '*B*' are two genes, '*ai*' and '*bi*' are the values of genes *A* and *B* for attribute '*i*', '$\bar{A}$' and '$\bar{B}$' are the respective mean values of *A* and *B*, 'c_A' and 'c_E' are the respective standard deviations of *A* and *B*.

In this work, we first divide the data-set, as we did in case of Euclidean distance, into number of sub data-sets. Then the Pearson co-efficient calculated between all the genes and stored in a matrix. Now 'del' value is as the minimum co-relation co-efficient, above which the genes are said to be highly positively co-related. Then it is similar to Euclidean distance, here the same method has been used to select the genes into our selected gene sub-set. The value of 'del' is calculated by hit and trial method.

2.3 Information Based Measures

Mutual information[15] is the measure of mutual dependence between two random variables in the case of probability theory, which is the base of information theory.

2.3.1 Dynamic Relevance

Dynamic relevance is an information theory based approach for gene selection using the concept of entropy. A step by step procedure is given below:

A selected subset (DRGS) is initialized to ∅. In the first phase the relevance of all the genes is calculated with the target class as:

$$R(g; class) = I(g; class) \quad (11)$$

The gene with the highest value of relevance is the first gene to be selected into our selected subset (DRGS). The Dynamic Relevance value (DR) of each gene is initialized to relevance value calculated in this step. The redundancy between the genes is calculated in the next phase. This is because we need to select those genes into our selected subset which are not redundant. Two genes can be said as redundant if their values are completely co-related[4]. In terms of information theory this can be expressed as:

$$I\left(g_i; class / g_j\right) \le I(g_i; class) \quad (12)$$

Genes are normally grouped as interdependent genes. Interdependence implies that one gene (g_i) needs the help of another (g_j) to do some functioning. And hence the relevance of gene (g_i) with the target class is incremented when conditioned with gene (g_j). In the third phase CR or C_Ratio is calculated for the selected gene (g_i) with the other genes (g_j). The gene with the highest value of CR is selected into the selected sub-set. In the next iteration CR is calculated between the newly selected gene and other genes. This process continues until the selected subset contains a user specified no of genes. The selection process is done keeping in mind that the CR value does not go into the negative side. Once it does so, selection is stopped there[4]. CR is calculated as:

$$CR_{i,j} = 2\frac{I\left(g_j; class / g_i\right) - I(g_j; class)}{H(g_j) + H(class)} \quad (13)$$

A positive value of CR indicates that the genes (g_i and g_j) are independent.

2.3.2 Information Gain

In using 'ID3' as a decision tree algorithm, it uses a attribute or feature selection measure called as information gain. In the year 1948, Claude Shannon developed the idea of information entropy which is the measure of uncertainty in a message and then further digging into it to lay the base knowledge for information theory where the information content of a message is calculated mathematically[13,16–18]. Using it in terms of a micro-array data-set and in decision trees, we know that decision tree is a tree where each node is actually a mini data-set where based on an attribute the data-set can be divided or classified. Now let we have a node 'N' where we hold the data-set or a part of it referred to as 'D'. Now here we calculate the information gain of all the attributes and the attribute which will have the highest information gain value is selected as the attribute for dividing the data-set into a number of partitions. This attribute is chosen as the splitting attribute. This information gain is calculated as in (13).

$$IG(t) = -\sum_{i=1}^{|c|}(p(c_i)\log p(c_i) + p(t))\sum_{i=1}^{|c|}\left(p\left(\frac{c_i}{t}\right)\log p\left(\frac{c_i}{t}\right) + p(\bar{t})\right) \times \sum_{i=1}^{|c|}\left(p\left(\frac{c_i}{\bar{t}}\right)\log p\left(\frac{c_i}{\bar{t}}\right)\right) \quad (14)$$

Where, c_i represents the i[th] gene, $p(c)_i$ represents the probability of that gene, 'p(t)' and 'p($\bar{t}$)' are the probabilities that the term 't' appears or not in the data-set, respectively.

The approach in this paper work is same as that of dynamic relevance. A selected subset 'IGGS' Information Gain Gene Selected is created with zero elements in it and a threshold value k is set as the maximum number of genes to be selected[15]. Then, with iteration a gene is selected. Thus the genes selected are highly co-related.

3. Schematic Representation of Proposed Model

Initially the data set has been pre-processed and then gene selection is applied in two different approaches. One is the information based approach and the other is statistical approach. The statistical approach includes Euclidian Distance (ED) and Pearson Co-efficient (PR) and information theory based approach includes Dynamic Relevance (DR) and Information Gain (IG). The generated subsets of interdependent genes by using ED and PR are then given as input to the learning machines where a sub-set is divided into training set and testing data-sets. First the train set is fed to the classifier followed by the test set where the classifier gives a class level to the test set which is compared with the original class level to calculate the accuracy. Finally a comparison is made between the outputs of the two classifiers. If required missing value imputation is done by *K*-Nearest neighbor method and Min–Max normalization is used to normalize the data-sets. Feature reduction is done through Principal Component Analysis (PCA). Two classifiers are used here, Naïve Bayes and Support Vector Machine (SVM). Naïve Bayes is a probabilistic classifier which is based on the Bayes' theorem. It is based on independent assumptions, that presence or absence of one gene does not affect the classification of another. SVM is a supervised classifier that tries to define boundaries between two or more classes by constructing a hyper plane or a set of hyper planes. SVM is meant for two class problems. So when we have three or more classes, it will work by considering two classes to be one and the third to be another class. This process goes on iteratively for all the classes.

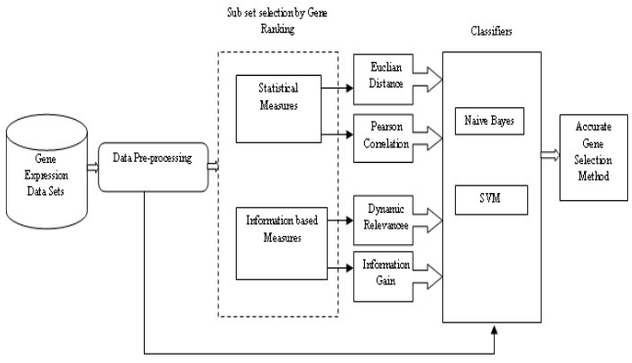

Figure 1. Schematic representation of proposed work.

4. Experimental Evaluation and Results

4.1 Data Set Description

The experimental evaluation has been conducted up on four publicly available dataset[20] data sets. A brief description of each data set[19,20] is shown in Table 1. The table includes name of the data set, number of genes, number of attributes, number of class and class detail.

Table 1. Data sets used in the gene selection experiments

Data Sets	No of Genes	No of attributes	No of class	Class Detail
Breast cancer	98	25	3	C1-11 C2-51 C3-36
Leukemia	72	256	2	Acute Lymphoblastic Leukemia-47 Acute Myeloid Leukemia: 25
Hepatitis	155	19	2	1. DIE-32 2. LIVE-123
Dermatology	366	34	6	Psoriasis-112 Seboreic dermatitis-61 Lichen planus-72 Pityriasis rosea-49 Cronic dermatitis-52 Pityriasis rubra pilaris-20

4.2 Pre-processing of Data Sets

As per the proposed work, the first step is pre-processing. The data set were normalized using *min-max normalization* technique. This technique makes all the data to fall within the user specified range of minim and maximum. Here, new minimum and new maximum are taken to be 0 and 1 respectively. Principal Component Analysis

(PCA)[13] has been used data reduction on the leukemia, hepatitis and dermatology datasets. The original and the reduced number of attributes are shown in Table 2.

Table 2. Data reduction

Data Sets	Original Size	Size after Reduction
Leukemia	72*256	72*165
Hepatitis	155*19	155*9
Dermatology	366*34	366*19

4.3 Gene Subset Selection

In the phase of gene sub-set selection, selection of subsets is done in the two different approaches discussed earlier i.e. mutual information approach and statistical approach. DR and IG are the mutual information based methods that have been used in this work. In the case of statistical approach one of the methods is Pearson co-efficient and the other is Euclidian Distance method. Using the mutual information approach, the subset is generated directly, but the statistical measures are methods of gene ranking. So from the ranked genes, the subsets are selected according to the threshold choices. The 'ED' method and the Pearson co-efficient method are a new approach in the field of gene selection as it was not applied earlier. Using each method, a number of subsets are produced according to the different threshold values and each subset generated is then applied on the classifier to generate the result which is discussed in the next phase.

4.4 Result Analysis

Before applying the classifiers on the selected data-set first it is required to split them into training and testing subset. Then the training subset along with its class levels is given as an input to our classifier, where it learns about the different feature patterns of the instances and the subsequent class to which it belongs. Then the testing set is supplied without showing the class level of that data-set. Here two classifiers are used, Naive Bayes and SVM. Naive Bayes is a probabilistic classifier which is based on the Bayes' theorem. It is based on independent assumptions, that presence or absence of one gene does not affect the classification of another. SVM is a supervised classifier that tries to define boundaries between two or more classes by constructing a hyper plane or a set of hyper planes. SVM is meant for two class problems. So when three are three or more classes, it will work by considering two classes to be one and the third to be another class. This process goes on iteratively for all the classes. The classifier then provides a class level to each instance of the testing data-set. This is compared with the original class level. Then the accuracy can be calculated in terms of correct classification. A conclusion based on the accuracy that which classifier acts better on which data-set has been explained in the Tables 3 and 4.

Table 3. Comparison of accuracy by using Naïve Bayes Classifier

Methods		Breast Cancer Acc (%)	Leukemia Acc(%)	Hepatitis Acc(%)	Dermatology Acc(%)
Original Data-Set		84.21	78.57	68.42	NA
Statistical Measure	Euclidean Distance	92.30	91.66	83.33	NA
	Pearson Co-efficient	96.55	91.66	88.45	NA
Information Measure	Dynamic Relevance	93.33	95.00	84.21	NA
	Information Gain	90.90	92.85	96.87	NA

Table 4. Comparison of accuracy by using SVM classifier

Methods		Breast Cancer Acc(%)	Leukemia Acc(%)	Hepatitis Acc(%)	Dermatology Acc(%)
Original Data-Set		36.84	64.28	57.89	32.46
Statistical Measure	Euclidean Distance	63.63	62.50	81.81	41.86
	Pearson Co-efficient	61.53	81.81	80.00	38.54
Information Measure	Dynamic Relevance	72.72	91.66	86.36	41.66
	Information Gain	58.82	71.42	82.14	34.06

5. Conclusion and Future Scope

Gene selection is an interesting approach for classification of diseases and is also applicable in the drug industry. The later is the case where the concept of functional group of genes is of utmost relevance, for this purpose interdependent gene which functions as a group is required to define functions of certain proteins. Dynamic relevance deals with this problem by integrating the advantages of mRMR approach with its interdependence calculation. Euclidian distance is a simple approach which takes into picture the

closeness of genes to find interdependent genes. In case of dynamic relevance, a set of interdependent genes are selected and it gives good accuracy results. In this paper, IG method shows better result when it is applied on Naive Bayes classifier where as DR method showing better result when it is applied on SVM classifier. In the paper[4] the authors devised DR method where a group of genes called as 'gene group' is selected, and this was applied on gene expression data-set with genes amounting to 20k. We have the intention of applying the same method on other data-sets and improve the same by applying some new method for gene ranking in the future. The output then can be validated with real-time medical records.

6. References

1. Kumar PG, Rathinaraja J, Victoire TAA. A combined MI-AVR approach for informative gene selection. Second International Conference on Sustainable Energy and Intelligent System (SEISCON); 2011. p. 870–5.
2. Fukuta K, Okada Y. Informative gene discovery in DNA microarray data using statistical approach. Intelligent control and innovative computing. Lecture notes in electrical engineering. 2012; 110:377–94.
3. Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, Solis DYW, Molter C, Duque R, Bersini H, Nowe A. GENESHIFT: A nonparametric approach for integrating microarray gene expression data based on the inner product as a distance measure between the distributions of genes. IEEE ACM Trans Comput Biol Bioinformatics. 2013 Mar/Apr; 10(2):383–92.
4. Sun X, Liu Y, Wei D, Xu M, Chen H, Han J. Selection of interdependent genes via dynamic relevance analysis for cancer diagnosis. J Biomed Informat. 2013; 46(2):252–8.
5. Khoshgoftaar TM, Wald R, Dittman DJ, Napolitano A. Comparing two new gene selection ensemble approaches with the commonly-used approach. 11th International Conference on Machine Learning and Applications (Volume:2); 2012. p. 184–91.
6. Sakar CO, Kursun O, Gurgen F. A feature selection method based on kernel canonical correlation analysis and the minimum redundancy–maximum relevance filter method. Expert Syst Appl. 2012; 39:3432–7.
7. Mandal M, Mukhopadhyay A. An improved minimum redundancy maximum relevance approach for feature selection in gene expression data. Procedia Technology. 2013; 10:20–7.
8. Chen Y, Zhang L, Li J, Shi Y. Domain driven two-phase feature selection method based on Bhattacharyya distance and kernel distance measurements. IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology; 2011. p. 217–20.
9. Ray SS, Bandyopadhyay S, Pal SK. New Distance measure for microarray gene expressions using linear dynamic range of photo multiplier tube. Proceedings of the International Conference on Computing: Theory and Applications; 2007. p. 337–41.
10. Piramuthu S. The Hausdroff distance measure for feature selection in learning applications. Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences, 1999. HICSS-32. (Volume:Track6); 1999.
11. Hsu HH, Lu MD. Feature selection for cancer classification on microarray expression data. Eighth International Conference on Intelligent Systems Design and Applications; IEEE; 2008. p. 153–8.
12. Chandra B, Gupta M. An efficient statistical feature selection approach for classification of gene expression data. J Biomed Informat. 2011; 44(4):529–35.
13. Han J, Kamber M. Data mining: concepts and techniques. 2nd ed. Morgan Kaufmann Publications; 2006.
14. Liu R, Yang N, Ding X, Ma L. An unsupervised feature selection algorithm: Laplacian score combined with distance-based entropy measure. Third international symposium on intelligent information technology application, 2009, IITA 2009; 2009 Nov; 65–8.
15. Cai R, Hao Z, Yang X, Wen W. An efficient gene selection algorithm based on mutual information. Neurocomputing. 2009; 72:991–9.
16. Uguz H. A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowl Base Syst. 2011; 24:1024–32.
17. Srimani PK, Koti MS. Knowledge discovery in medical data by using rough set rule induction algorithms. Indian Journal of Science and Technology. 2014 Jul; 7(7):905–15.
18. Rajkumari SB, Nalini C. An efficient data mining dataset preparation using aggregation in relational database. Indian Journal of Science and Technology. 2014 Jun; 7(S5):44–6.
19. UCI machine learning repository. 2012. Available from: http://www.ics.uci.edu/~mlearn/MLRepository.html
20. Dermatology Data Set. 2014 Apr 18. Available from: http ://archive:ics:uci:edu/ml/datasets/Dermatology