Using Feature Minimization Technique to Resolve Minimum Constraint Cover Problem.

Author-Madhugandha Bhosale

Department of Computer Science, BAMU University,

City-Aurangabad, State-Maharashtra, Country-India,

Abstract -
The amount of internet data has grown a few years ago. And cause the risk of data problems in accessing information related to users Online user data requirements can be calculated by evaluating the web navigation behavior of users. Web Usage Mining (WUM) is used to extract knowledge from the user's web access logs using the Data Mining technique, one of the applications. WUM's steepness is a website referral system, which is a technique for filtering personal information used to determine whether some users will approve a given item or to identify items that can provide The major users In this article, a modified architecture that combines item data with user access log data and then searches for patterns and creates pattern groupings. After that, create a set of instructions for users. Therefore the execution time and the run time is reduced.Other experiments compare CFS with a wrapper, which is a well-known method of selecting features that use the target learning algorithm to evaluate the set of properties. In many cases, CFS gives results that are equivalent to envelopes and typically envelopes That is smaller than the CFS data set. It works faster than a wrapper, making it possible to expand to a larger data set.

Keywords: *Feature selection, feature ranking, redundancy minimization, Radial Basis Function,Kernel*

Introduction

Digging on the web is the use of data mining systems to focus on learning from information on the web, including archiving on the web, hyperlinks between reports, using website logs, and more. Can be categorized into three categories without distortion as specified in the data to be excavated. There are three types of web mining that we mentioned below. Web content mining is a way to extract valuable information from the content of a web data warehouse. Content information is a collection of confidence that is available on the website. It may include content, images, audio, video or organizing, for example, notes and tables. The use of digging content for web content has been reviewed as much as possible. Problems that often arise in content digging include point disclosure and tracking, deletion, linking, design, grouping of web archives and arrangement of website pages. Mining Web structure The structure of the web mill chart consists of a webpage, a hub, and a hyperlink that is connected to the relevant page. The Web Structure Mining is a method for finding structural information from the web. This can be separated into two categories according to the structure data used..

There are generally three forms of qualitative selection methods in literature: (1) filtering methods [14] by choosing independently from the classifier (2) wrapping method [12] which uses a black box prediction method to Sub-scores of qualifications and (3) integrated methods in which the selection process is integrated into the training process directly In the bioinformatics application, there are many applications and methods for these categories.

The most widely used filter method is F [4], relaxation [11, 13], mRMR [19], T test and data acquisition [21] which calculates the relationship sensitivity or Relevance) of the characteristics associated with wrt) distribution of data class labels These methods can be determined by using statistical data around the world. The method of selecting the type of wrapper is strongly associated with a specific classifier, such as selecting the relative properties (CFS) [9]. Support vectors.

Recursive Eliminator (SVM-RFE) [8] They often work well. But their calculation cost is very expensive. Recently, the uniformity of dimension reduction has been studied extensively, sparsity and used in the study. Selection of 1-SVM characteristics is proposed to be selected. Characteristics using the standard definition 1 normal, which tends to provide a dispersed solution [3] .Because the number of functions selected using SVM-1 is larger than the sample size, therefore M is proposed VS Huberized Hybrid MVS (HHSVM) including standard 1 and standard 2 to create more structured regulations .But it is designed for binary identification only In multi-task learning in parallel work,

Et al., [18] and Argyriou et al. Al [1] developed a similar model for standardization of 2.1 to select characteristics between tasks Such regulation has a close relationship with the snare group. [28] In this article, we present a new, effective and efficient method of selecting characteristics to use to reduce the sharing of norms 2.1 in functions. Loss and normalization Instead of using the standard loss function 2 that is sensitive to abnormal values, the loss-based functions that are standard 2.1 will be used in our work to suppress the use of unusual values. Inspired by previous research [1, 18], the normal '2.1 normal' operation will be performed to select the characteristics of all data points with common sparsity, such as each characteristic. (Gene expression or mass score value for all data scores or all large scores Data points To solve the new purpose of choosing effective features, we offer an efficient algorithm to solve this problem of normative reduction. Our algorithm Our method is better than five commonly used character selection methods in statistical and bioinformatics learning.

An Efficient Algorithm

Data matrix has been processed in advance and fragmented, taking into account the meaning of each expression of the gene (column), the number of output properties (genes), suggesting that n is provided from outside by the user. Data matrix with class c = {1, 2, •••, C} is the input at first. The first objective (obj1) is that the relevance of each gene is calculated by sharing data according to equation 6 from the score The relevance will be extracted and added to the identity of the top scorer.

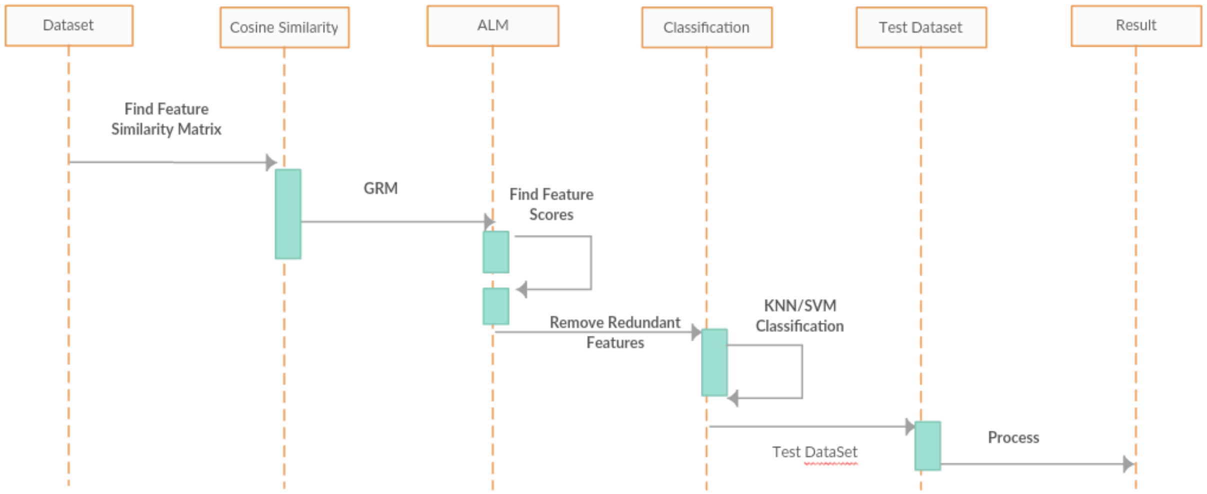

Figure 1.0 Sequence for proposed architecture

Algorithm 1 Proposed Feature Selection

Input: Dataset having many features
Output: Reduction in number of features
1 .Load dataset having many attributes or features and standard parameters.

2. Calculate Vertical aggregation of individual columns called weight.

3.That all aggregation values are compare with each other single column aggregation value.

4.Which columns difference is less take it first (similarity) .

5.Check by discarding one column & check for classification if classification is right then column can be discarded. If classification is wrong then that column can not be discarded .

6.Then go for next .

7.End

in the final solution set. Next a looping is performed for the remaining output features. Now the redundancy between the output feature and the remaining features (*idle f t*) is calculated as per Equation 5. If the output feature set contains more than one feature then the mean is considered as the redundancy score as in Equation .

$$\text{mean-redundancy}(i) = \sum_{k=1}^{F} (\text{mutual-info}[x_k, x_i]))/|F|),$$

where *F* is output feature set, *Xk* is output feature vector and *xi* is the *i*th feature vector. Then the second objective is modeled as the ratio of relevance to the redundancy and it is to be maximized. After calculating the two objectives for each feature the non-dominated features are identified. A reference feature is called the non-dominated feature if it satisfies the following conditions: 1) if the obj1 of the reference feature is greater than or equal to all the other futures' obj1 and the obj2 of the reference feature is greater than or equal to all the other features' obj2 2) if the obj1 of the reference feature is greater than all the other features' obj1 and the obj2 of the reference feature is less than all the other features' obj2 and vice-versa. Afterwards, from the non-dominated features, the feature having maximum obj2 is included in the output feature set.

Datasets and Results

Four real life datasets are used for comparative study. One real life data sets is used for the comparative study. The Pima Indian Diabetes dataset is collected from the most popular UCI Repository. The dataset contain two classes of samples.

1. 1. Risk Factor of Cervical cancer: This standard dataset is taken from most popular repository that is UCI Repository.
2. Risk Factor of Cervical cancer (Reduced version): This dataset is taken from kaggle.com, which is reduced dataset of Risk factor of cervical cancer.
3. Pima Indian Diabetes : This standard dataset is taken from most popular repository that is UCI Repository.
4. Pima Indian Diabetes (Reduced version): This dataset is taken from kaggle.com, which is reduced dataset of Pima Indian Diabetis.

Risk_Factor of cervical cancer

Datasets	Precision	Recall	Accuracy	F-measure
UCI Repository (all values in %)	93.33	93.33	90.90	46.60
Kaggle.com (Reduced dataset) (all values in %)	93.75	93.75	93.103	46.87

Table 1.Comparitive Result

Pima Indian Diabetes

Datasets	Precision	Recall	Accuracy	F-measure
UCI Repository (all values in %)	80	80	71.42	40
Kaggle.com (Reduced dataset) (all values in %)	60	75	57.142	33.33

Table2.Comparitive Result

(1) Our comparison go through four datasets in which we taken Cervical cancer dataset which is standard and got from UCI Repository. And Reduced dataset got from Kaggle.com for same standard dataset. Both having higher number of attributes get more accuracy results.

(2) Then we through next two datasets which having less number of attributes. We taken Pima Indian Diabetes dataset .And Reduced dataset got from Kaggle.com for same standard dataset. We get less accuracy when number of attributes/features are less.

(3) So, our limitation is that when number of attributes/features are less our system is not work that much superior .

(4) Our application is that our model work good in prediction model mostly in real life area like prediction of diseases when there are higher number of attributes/features.

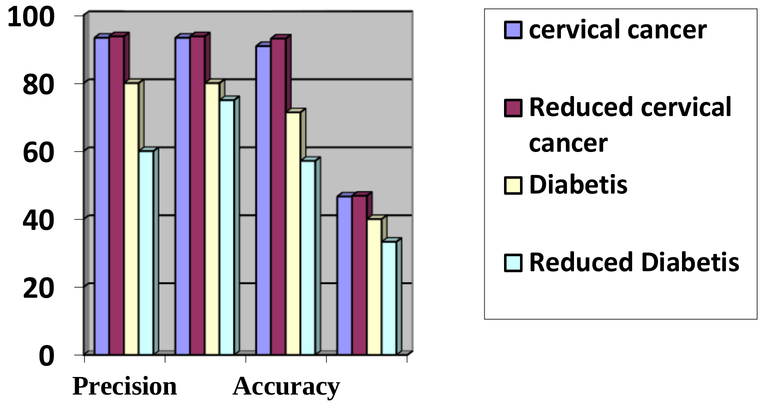

Figure 8: Graphical Results

Conclusion

There are different types of feature selection methods available in the existing literature. But in most cases, we have seen that the fundamental objective of the method is either relevance or redundancy. In this paper, we have proposed a method where relevance and redundancy are supported in parallel. Redundancy is described as mutual information among the characteristics. The number of resulting functions is provided by the user. The performance of the proposed technique is evaluated on the basis of some sets of real life microarray gene expression data to select non-redundant and relevant genes. In addition, the performance of the proposed method is compared with different standard datasets which having their own accuracy.
As we got comparative results from our system. We got some good results and bad results. In good results we get more accuracy .In bad results we get less accuracy . So I conclude that our system is work superiorly with higher number of attributes and not that much superior with less number of attributes.

References

[1] Pena, J.M., Lozano, J.A., Larranaga, P., Inza., I.. Dimensionality reduction in unsupervised learning of conditional gaussian networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 2001;23(6):590–603.

[2] Kurun, O., Akar, C.O., Favorov, O., Aydin, N., Urgen, F.. Using covariates for improving the minimum redundancy maximum relevance feature selection method. Turkish Journal of Electrical Engineering and Computer Sciences 2010;18(6):975–987.

[3] Kamandar, M., Ghassemian, H.. Maximum relevance, minimum redundancy band selection for hyperspectral images. In: 19th Iranian Conference on Electrical Engineering (ICEE),. 2011,.

[4] Dy, J.G., Brodley, C.E., Kak, A., Broderick, L.S., Aisen, A.M.. Unsupervised feature selection applied to content-based retrieval of lung images. IEEE Transaction on Pattern Analysis and Machine Intellegence 2003;25(3):373–378.

[5] Zhang, Z., R.Hancock, E.. A graph-based approach to feature selection. In: International Workshop on Graph-Based Representations in Pattern Recognition. 2011,.

[6] Cai, D., Zhang, C., He, X.. Unsupervised feature selection for multi-cluster data. In: 16th ACM SIGKDD International Conference on Knowledge Discovery and Data mining. 2010,.

[7] Ruiza, R., Riquelmea, J.C., Aguilar-Ruizb, J.S.. Incremental wrapper-based gene selection from microarray data for cancer classification. Pattern Recognition 2006;39(12):2383–2392.

[8] Mitra, P., Murthy, C., Pal, S.K.. Unsupervised feature selection using feature similarity. IEEE Transaction on Pattern Analysis and Machine Intellegence 2002;24(3):301–312.

[9] Sondberg-Madsen, N., Thomsen, C., Pena, J.M.. Unsupervised feature subset selection. In: In Proc. of the Workshop on Probabilistic Graphical Models for Classification. 2003,.

[10] Ding, C.H.Q.. Unsupervised feature selection via two-way ordering in gene expression analysis. Bioinformatics 2003;19(10):1259–1266.

[11] Kohavi, R., John., G.. Wrapper for feature subset selection. Artificial Intelligence 1997;97:273–324.

[12] Jiang, S., Wang, L.. An unsupervised feature selection framework based on clustering. In: New Frontiers in Applied Data Mining. 2008,.

[13] Morita, M., Oliveira, L.S., Sabourin, R.. Unsupervised feature selection for ensemble of classifiers. In: Frontiers in Handwriting Recognition. 2004,.

[14] Handl, J., Knowles, J.. Feature subset selection in unsupervised learning via multiobjective optimization. International Journal of Computational Intelligence Research 2006;2(3):217–238.

[15] Dash, M., Liu, H.. Unsupervised feature selection. In: In Proc. Pacific Asia Conf. Knowledge Discovery and Data Mining. 2000,.