Efficient Feature Selection by using Global Redundancy Minimization and Constraint Score

Akansha A. Tandon, Sujata Tuppad

Abstract*: *A central problem in automatic learning is the identification of a representative set of characteristics from which to construct a classification model for a particular task. This thesis deals with the problem of the selection of characteristics for automatic learning by a correlation - based approach. The central assumption is that good sets of characteristics contain characteristics that are highly correlated with the class but not correlated with each other. A formula for evaluating characteristics, based on ideas derived from test theory, provides an operational definition of this hypothesis. CFS (Correlation based Feature Selection) is an algorithm that couples this evaluation formula with an appropriate correlation measure and a heuristic search strategy. Other experiments compared the CFS to a wrapper - a well-known approach to feature selection that uses the target learning algorithm to evaluate sets of features. In many cases CFS has given results comparable to the envelope, and in general, surpassed the envelope on small sets of data. CFS runs much faster than the wrapper, enabling it to extend to larger sets of data.

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

I. INTRODUCTION

The selection of functions, the process of selecting a subset of relevant functions, is a key element in the construction of robust automatic learning models for classification, clustering and other tasks. The functions section plays an important role in many applications, as it can accelerate the learning process, improve the generalization capability of modes and mitigate the curse's effect of dimensionality[15].In past there are large number of developments on the selection of characteristics that have been made in the literature and there are many recent reviews and workshops devoted to this subject, for example, NIPS Conference [7]. Over the last ten years, the selection of characteristics has seen many activities mainly due to advances in bioinformatics where a large amount of genomic and proteomic data are produced for biological and biomedical studies. For example, in genomics, DNA microarray data measure the expression levels of thousands of genes in a single experiment. Gene expression data usually contain a large number of genes, but a small number of samples. A given disease or biological function is usually associated with a few genes [18]. Over several thousand genes to select some of the relevant genes thus becomes a key problem in bioinformatics research [19]. In proteomics, the high-throughput mass spectrometer (MS) measures the molecular weight of individual biomolecules (such as proteins and nucleic acids) and has the potential to discover putative proteomic biomarkers. Each spectrum is composed of peak amplitude measurements at about 15,500 characteristics represented by a corresponding load mass value. The identification of significant proteomic characteristics of MS is crucial for disease diagnosis and profiling of protein biomarkers [19].

In general, there are three models of characteristic selection methods in the literature: (1) filtering methods [14] where selection is independent of classifiers, (2) wrapping methods [12] where the method of Prediction is used as a black box to score subsets of features, and (3) integrated methods where the feature selection procedure is integrated directly into the training process. In bioinformatics applications, many methods for selecting the characteristics of these categories have been proposed and applied.

Methods for selecting widely used filter characteristics include statistical F [4], relief [11, 13], mRMR [19], t-test and information gain [18] which calculate sensitivity Correlation, or relevance) of a characteristic with respect to (wrt) the class label distribution of the data. These methods can be characterized by the use of global statistical information. Wrapper type selection methods are tightly coupled to a specific classifier, such as the correlation-based feature selection (CFS) [9], the support vector

Recursive elimination machine (SVM-RFE) [8]. They often perform well, but their computational cost is very expensive. Recently, the regularity of sparsity in the reduction of dimensionality has been widely studied and also applied in characteristic selection studies. 1-SVM was proposed to perform characteristic selection using 1-normal regularization which tends to give a scattered solution [3]. Because the number of selected functionalities using SVM-1 is greater than the sample size, a Huberized Hybrid MVS (HHSVM) was proposed combining both Standard 1 and Standard 2 to form a more structured regularization. But it was designed only for binary classification. In multi-task learning, in parallel work, Obozinsky Et al., [18] and Argyriou et al. Al. [1] developed a similar model for the regularization of the 2.1 standard to couple the selection of characteristics between tasks. Such regularization has close ties with the group lasso [28]. In this article we propose a new efficient and robust method of characteristic selection to use the joint minimization of the norm 2.1 on the loss function and the regularization. Instead of using a loss function based on standard 2 that is sensitive to outliers, a loss function based on the 2.1 standard is adopted in our work to suppress outliers. Motivated by previous research [1, 18], a '2.1 normal'

Revised Version Manuscript Received on December 14, 2016.

Akansha A. Tandon, Department of Computer Science & Engineering, BAMU Matsyodari Shikshan Sansthas College of Engineering and Technology Jalna, Aurangabad (Maharashtra)-431203. India.

Sujata Tuppad, Assistant Professor, Matsyodari Shikshan Sanstha's College of Engineering and Technology, Jalna, Aurangabad (Maharashtra)-431203. India.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

regularization is performed to select characteristics across all data points with common sparsity, ie each characteristic (expression Gene or mass-Scores value for all data points or has large scores on all data points To solve this new objective of robust characteristic selection we propose an efficient algorithm to solve this problem of minimization of the norm 2.1 We also provide algorithmic analysis and prove the convergence of our algorithm. We have extensive experiments on six sets of bioinformatics data and our method outperforms five other commonly used methods of character selection in statistical and bioinformatics learning.

II. AN EFFICIENT ALGORITHM

The data matrix has been preprocessed and discretized with respect to the mean of each gene's expression (column). The number of output features (genes) say n is provided from outside by the user. The data matrix with classes $c = \{1, 2, \cdots, C\}$ are the inputs. At the beginning, the first objective (obj1) i.e., the relevance of each gene is calculated by mutual information as per Equation 6. From the relevance score, the highest scorer gene id is extracted and added.

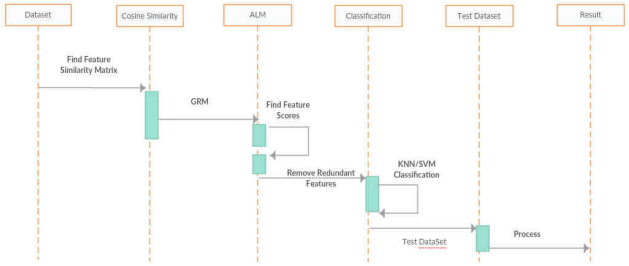

Figure 1.0 Sequence for proposed architecture

Algorithm 1. Proposed Feature Selection

Input: The feature id *idle f t*, first objective *ob j*1, second objective *ob j*2, $|ob\ j1| = |ob\ j2| = |idle\ f\ t|$.

Output: Non-dominated feature id *idns*, the second objective *ob j*2*ns* of non-dominated features.

1: $k = 1$;
2: for $i = 1 : |idle\ f\ t|$ do
3: $t = 0$;
4: for $j = 1 : |idle\ f\ t|$ do
5: if then$(i! = j)$
6: if then$(ob\ j1(i) \leq ob\ j1(j) \& ob\ j2(i) \leq ob\ j2(j))$;
7: else if then$(ob\ j1(i) < ob\ j1(j) \& ob\ j2(i) > ob\ j2(j) \| ob\ j1(i) > ob\ j1(j) \& ob\ j2(i) < ob\ j2(j))$;
8: else
9: $t = 1$;
10: break;
11: end if
12: end if
13: end for
14: if then$(t == 0 \& j == |idle\ f\ t|)$
15: $idns(k) = i$;
16: $ob\ j2ns(k) = ob\ j2(i)$;
17: $k = k + 1$;
18: end if
19: end for

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 9.

$$\text{mean-redundancy}(i) = \sum_{k=1}^{F} (\text{mutual-info}[x_k, x_i])/|F|), \quad \ldots \text{Eq. 9}$$

where F is output feature set, Xk is output feature vector and xi is the ith feature vector. Then the second objective (obj2) 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.

III. DATASETS & RESULTS

One real life data sets is used for the comparative study. The Prostate Cancer dataset is collected from the website: www.biolab.si/supp/bi-cancer/projections/info/. The dataset contain two classes of samples.

1. Prostate: Gene expression measurements for samples of prostate tumors and adjacent prostate tissue not containing tumor were used to build this classification model. It contains 50 normal tissue and 52 prostate tumor sample. The expression matrix consists of 12533 numbers of genes and 102 numbers of samples.

Table 1.0 Result with Existing Methods

Data Set	method	Sensitivity	Specificityy	Accuracy	Fscore	Avg Corr	AUC
Prostate Cancer	**Proposed method**	0.98	0.9423	0.9608	0.9608	0.23	0.9892
	mRMR (MID)	0.96	0.9038	0.9314	0.932	0.322	0.9592
	mRMR (MIQ)	0.978	0.923	0.951	0.9513	0.237	0.983

The actual data sets described above are first standardized with the Min-Max normalization technique. Then, with respect to the mean of each characteristic (gene) or column, the data are discretized. In this article, the number of output functions is taken as 100 for all algorithms. Using 10-fold cross-validation, sensitivity, specificity, precision and fscore score are calculated. Then, the mean correlation for evaluating the redundancy of the selected characteristics is also calculated. A smaller correlation value indicates that the selected functions are less redundant. In addition, the area under ROC curve (AUC) is also reported.

The metric performance values of the proposed method, mRMR (MID) and mRMR (MIQ) on the different real datasets are shown in Table 1. It is evident from the table that for the data series on cancer Prostate sensitivity, specificity, and AUC are respectively 0.98, 0.9423, 0.9608, 0.9608 and 0.9892, which are better than the mRMR (MID) and mRMR (MIQ) patterns in all cases. Furthermore, the average correlation of the proposed method is 0.23, which is lower than the other two methods and indicates that the resulting

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

characteristics given by the proposed method are the least correlated.

Summary Results

Iterations	3595
Total basis functions used	57
Number correct	26
Number incorrect	8
Percentage correct	76

Iterations	7948
Total basis functions used	701
Number correct	26
Number incorrect	8
Percentage correct	91.17

Confusion matrix:

	Normal	Tumor
Normal	9	0
Tumor	8	17

Confusion matrix:

	Normal	Tumor
Normal	9	0
Tumor	3	22

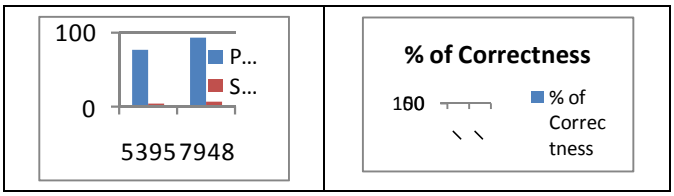

Figure 1.1 Graph for Categorization of True Positive

IV. 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. To measure the relevance and redundancy of a characteristic or a gene, mutual information was considered. Relevance is defined as the mutual information between a feature vector and class labels. 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 that of well-known mRMR (MID), and mRMR (MIQ) and the results show that the proposed method over performs mRMR schemas for all data sets.

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
16. P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, “Eigenfaces vs. fisherfaces: Recognition using class specific linear projection,”IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, no. 7, pp. 711–720, Jul. 1997
17. X. Chang, F. Nie, Y. Yang, and H. Huang, “A convex formulation for semi-supervised multi-label feature selection,” in Proc. AAAI Conf. Artif. Intell., 2014, pp. 1171– 1177
18. J. Wu and J. M. Rehg, “CENTRIST: A visual descriptor for scene categorization,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 33,no. 8, pp. 1489–1501, Aug. 2011 [5] G. Forman and E. Kirshenbaum, “Extremely fast text feature extraction for classification and indexing,” in Proc. Int. Conf. Inf. Knowl. Manag., 2008, pp. 1221–1230
19. R. Kohavi and G. H. John, “Wrappers for feature subset selection,” Artif. Intell., vol. 97, no. 1/2, pp. 273–324, 1997

Akansha A. Tandon, is pursuing her Masters in Engineering from MSSCET Jalna. Her hobbies include reading books and listening music. Special Thanks to Principal Dr C.M. Sedani

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.