Asian Journal of Computer Science and Technology
ISSN: 2249-0701 (P) Vol.7 No.S1, 2018, pp.108-111

DOI: https://doi.org/10.51983/ajcst-2018.7.S1.1793

Selection of Features on Mining Techniques for Classification

Gurrampally Kumar[1], S. Mohan[2] and G. Prabakaran[3]
[1]Research Scholar, [2&3]Assistant Professor
[1,2&3]Department of Computer Science and Engineering, Annamalai University, Tamil Nadu, India
E-Mail: grk.040@gmail.com, mohancseau@gmail.com, gpaucse@yahoo.com
(Received 5 October 2018; Revised 20 October 2018; Accepted 2 November 2018; Available online 9 November 2018)

***Abstract* - Feature selection has been developed by several mining techniques for classification. Some existing approaches couldn't remove the irrelevant data from dataset for class. Thus it needs the selection of appropriate features that emphasize its role in classification. For this it consider the statistical method like correlation coefficient to identify the features from feature set whose data are very important for existing classes. The several methods such as Gaussian process, linear regression and Euclidean distance have taken into consideration for clarity of classification. The experimental results reveal that the proposed method identifies the exact relevant features for several classes.**
Keywords: **Feature Selection, Data Mining, Classification, Correlation Coefficient**

I. INTRODUCTION

The feature selection has been issued in data mining based on different methodologies to reduce and control the dimensions. The correlated features are emphasized for classification by several selecting approaches in data mining. Sometimes the feature selection or identification of features is not perpetrated for classification. Apart from several performances on feature selection and its analysis, the classification can be also improved by using several statistical methods. The several classification approaches are explained by different authors and researchers like the classification has taken on gene expression analysis that involves large number of features and micro array based classification [1]. The feature selection for biological data classification has been considered by [2, 3, 4, 5, 6, 7] for generating diagnostic classification systems. The optimization model [8] has also taken for classification in other direction. Recently another way of feature selection has been developed by maximizing independent classification information [9].

Further the new feature can be included for analyzing the classification problem, but the concept of tiny feature data from each feature is considered for generating new class. Initially, Bhuyan and Kamila have initiated to design the concept of sub-feature data and applied in [10, 11] using different database. Although they have used the sub-feature selection data as their own model, still it needs to develop different sub-feature selection model for classification. In this paper the correlation framework has taken to analyze the redundancy of features data based on mathematical model. The frameworks have been performed the well experimental evaluation based on proposed work. The experimental results have been generated as per the proposed model. The related works have described in next section.

II. RELATED WORK

Since this paper involves with feature selection as well as identification of features, the background of this research work is considered based on two parts. In first part, the related feature selection approaches are described whereas sub-feature selection based on respective selection approaches are elaborated in second part. The second part is very less due to lack of inadequate information regarding the corresponding related research work.

A. Several Approaches of Feature Selection

The several feature selection approaches have been developed by different researchers as in [12, 13, 14, 15, 16, 17, 21]. The unsupervised approaches have been developed by most of the researchers for feature selection method. Mitra *et al.,* [15] is proposed the partition of feature set into number of clusters using unsupervised feature selection scheme for reducing feature set. They observed the maximum information compression index approach is determined over various data sizes of real life data set. He *et al.,* [18] describes that each feature evaluation is determined by the power of locality preserving or Laplacian score while the graph Laplacian for feature selection has described in [19] for multi-cluster data. Sigmoid function has considered for clarification of correlation among features. Recently Bhuyan and Reddy [22] have developed feature selection based on correlation coefficient.

B. Approaches of Sub-Feature Selection

Since few papers on sub-feature selection are available in different sites, it only considers the concept of sub-features that explained in [10, 11] and helped to design the model of sub-feature selection for the proposed work. Next, it considers the framework for correlation among features for classification.

III. PROBLEM STATEMENT

The researchers have been focused on feature selection to remove and control the irrelevant features in a database for better classification performance. Among them, Banerjee

et al., [21] have shown that how the irrelevant features are discarded and controlled the redundancy from any data set. But when this model is considered for classification the corresponding method may not be regulated for predicted class when less number of features are available in data sets because above method selects very few features or no feature in some cases. But for large number of features in data sets, this method may consider in certain cases.

Several approaches can be used to measure the appropriate feature selection based on class labels. If class labels are not known, Sammon's error can be used for formulating of the feature extraction. Let X $\subseteq R^{n\times p}$ be the data set i.e., X has n data samples and p dimensions (or features) X can be represented by X = {x_1, x_2, …. x_n} where x_i = (x_{i1}, x_{i2}, …. x_{in})^T, i=1…..n. The Euclidean distance between x_i, x_j represent as

$$d_{ij}^{o}(x_i, x_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \quad (1)$$

Then the sammon's error can be formulated based on above Euclidian distance as follows

$$SE = \frac{1}{\sum_{i<j} d_{ij}^{o}} \sum \frac{(d_{ij}^{o} - d_{ij}^{r})^2}{d_{ij}^{o}} \quad (2)$$

Where $d_{ij}^{r}(x_i, x_j)$ is the reduced dimensional distance

To avoid dependent features, it can consider the correlation coefficient between ith & jth feature and control the redundancy. The non-linear correlation is considered for avoiding constant ratio to the amount of change in the other features values.

IV. CORRELATION FRAMEWORK ON FEATURE SET

To avoid irrelevant dependent features and control redundancy, it generates a method to assess the redundancy among the selected features based on correlation coefficient between ith & jth features. Thus the equation can be written as

$$CR = \frac{1}{(p-r)} \sum_{i=1}^{p} w_i \sum_{j\neq i} \rho_{ij} w_j \quad \text{with } p \neq r \quad (3)$$

Where CR – controlling redundancy, w_i & w_j weight factor of ith & jth feature, ρ_{ij} - correlation coefficient between ith & jth features, p – total no. of features, r – selected features. Although the activation function has been defined as the definition 1, but based on that activation function, the sigmoid function can be defined accordingly [20].

Definition 1: A measurable function S: R → R is called "activation function" whenever

$$S(x) = a \text{ and } S(x) = b \quad \text{with } a \neq b. \quad (4)$$

These functions must have a graph with the same behavior of the unit step function and these leads to the introduction of a new class of functions called the sigmoidal function which is defined as follows.

Definition 2: A measurable function σ: R→R is called "a sigmoidal function" whenever

$$S(x) = a \text{ and } S(x) = b \quad (5)$$

These functions are smooth and bounded sigmoidal function. Further Benjamin Gompertz introduced smooth sigmoidal function that can be generated as follows

$$\sigma_{\alpha,\beta}(x) = e^{-\alpha e^{-\beta x}}, \quad x \in R \quad (6)$$

Where $\alpha, \beta > 0$ represent an effective translation and scaling term respectively.

It considers the smooth and bounded function, specially, Gompertz function for proposed model. Thus the equation 3 can be rewritten as

$$CR = CR = \frac{1}{(p-r)} \sum_{i=1}^{p} \sigma_{\alpha,\beta}(x_i) \sum_{j\neq i} \rho_{ij} \sigma_{\alpha,\beta}(x_j), \quad \text{with } p \neq r \quad (7)$$

Where $w_i = \sigma_{\alpha,\beta}(x_i)$ and $w_j = \sigma_{\alpha,\beta}(x_j)$

For above equations are considered to control redundancy based on reduced features whereas it is more comfortable to select sub-feature for unique class.

V. EXPERIMENTS

This section explains the experimental evaluation based on different dataset. It considers only two data sets i.e., abalone and arrhythmia. The performance of proposed method is explained based on the above data sets. The data set collect from UCI machine learning repository.

TABLE I COVARIANCE MATRIX WITH HIGHEST AND LOWEST VALUE

Covariance term	Abalone	Arrhythmia
Average Target Value	0.2048611111111111	0.3778467908902686
Inverted Covariance Matrix		
Lowest Value = -	- 0.15726553207753707	-0.24917347888095953
Highest Value =	0.9340805197478603	0.8457119920304339
Inverted Covariance Matrix * Target-value Vector:		
Lowest Value =	-0.2598660289664312	-0.48841990864418083
Highest Value =	0.7499559617447284	0.5709794023348639

For experiment, it considers 10 fold cross validation for Gaussian processes, Linear regression and Euclidean distance. The covariance matrix has evaluated with highest and lowest value in Gaussian processes where the average

target value and its multiplication with covariance matrix as shown in table I. When the average target value increase among above two data set, it observed that lowest value increased where highest value decreased.

The cross validation is generated of the above data set that has mentioned in table II, III, and IV based on Gaussian processes, Linear regression and Euclidean distance. When it considers cross validation for linear regression model, its statistical value is changed as compare to Gaussian processes that is mentioned in table III.

TABLE II GAUSSIAN PROCESSES BASED CROSS VALIDATION RESULT

Cross Validation Term	Abalone	Arrhythmia
Correlation coefficient	-0.2763	0.2135
Mean absolute error	1.6605	3.9584
Root mean squared error	2.1571	5.2765
Relative absolute error	106.9615 %	99.3409 %
Root relative squared error	103.767 %	114.6309 %

TABLE III LINEAR REGRESSION BASED CROSS VALIDATION RESULT

Cross Validation Term	Abalone	Arrhythmia
Correlation coefficient	-0.2529	0.0285
Mean absolute error	1.8896	6.3383
Root mean squared error	2.3968	10.1954
Relative absolute error	121.7166 %	159.0684 %
Root relative squared error	115.298 %	221.494 %

When it considers nearest neighbor(s) for classification, its values have been changed as shown in table IV. From table II, III and IV, it observed that the correlation coefficient is decreased, but the values of other terms are increased in dataset abalone. In dataset Arrhythmia, the values of different terms are not appropriate as per the order of terminology such as Gaussian processes, linear regression and Euclidean distance. Thus in equation 7, when the number of selected feature is increased, the correlation coefficient values are increased. The variable p and r are choosing very cautiously to find selected features. If the more number of features will be selected, there is no effect of classification. It is very difficult to find feature to identify distinguished class among classes in particular dataset.

TABLE IV EUCLIDEAN DISTANCE BASED CROSS VALIDATION RESULT

Cross Validation Term	Abalone	Arrhythmia
Correlation coefficient	-0.1209	0.1553
Mean absolute error	2.1389	3.7633
Root mean squared error	2.9011	5.6641
Relative absolute error	137.7749 %	94.4443 %
Root relative squared error	139.5588 %	123.0523 %

From above techniques, Euclidean distance is better than other techniques for dataset abalone, but in case of Arrhythmia dataset, Gaussian processes is better than other techniques. Thus the considered techniques are more appropriate based on the number of selected features.

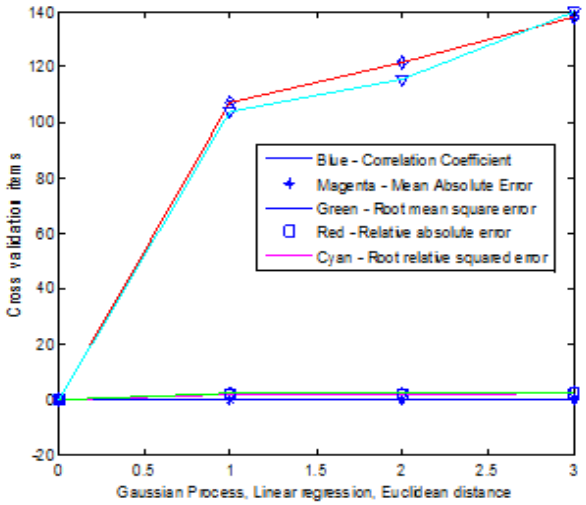

Fig. 1 Distinguished Cross validation among Gaussian process, linear regression and Euclidean distance on Abalone dataset

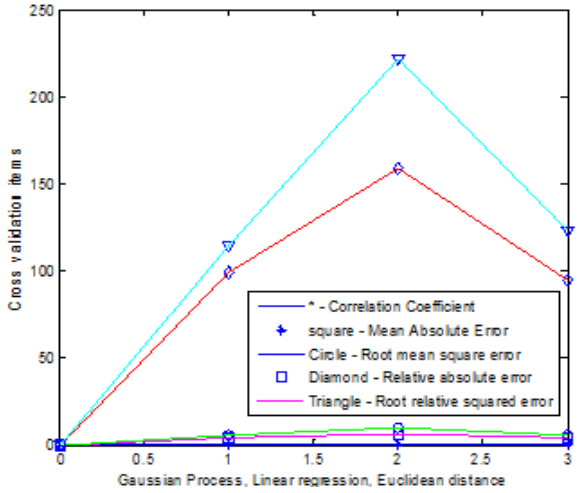

Fig. 2 Distinguished Cross validation among Gaussian process, linear regression and Euclidean distance on Arrhythmia dataset

The feature data are evaluated through correlation coefficient and it is shown that the items of cross validation matrix are very close to each other in figure 1 and 2. For feature selection or identification of features, it can be considered Gaussian process and Euclidean distance methods for classification.

VI. CONCLUSION

In this paper, it has been proposed the model for feature selection based on correlation coefficient model. It is evaluated by Gaussian process, linear regression and Euclidean distance for Cross validation for feature identification. From above method, two method such as Gaussian process and Euclidean distance methods are very effective for classification. Although these methods have

considered for classification, but the tiny data of each feature is very important to generate distinguished class among different classes in a dataset which is the future work.

REFERENCES

[1] E. R. Dougherty, "Small sample issue for Microarray-based classification," *Comparative Functional Genomics*, Vol. 2, pp. 28–34, 2001.

[2] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *in Proc. Compute. Syst. Bioinformatics Conf*., pp. 523–529, 2003.

[3] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, Vol. 286, pp. 531–537, Oct. 1999.

[4] N. R. Pal, K. Aguan, A. Sharma, and S. Amari, "Discovering biomarkers from gene expression data for predicting cancer subgroups using neural networks and relational fuzzy clustering," *BMC Bioinformatics*, Vol. 8, pp.1-18, 2007.

[5] N. R. Pal, "A fuzzy rule based approach to identify biomarkers for diagnostic classification of cancers," *in Proc. IEEE Int. Fuzzy Syst. Conf*., pp. 1–6, 2007.

[6] Y.-S. Tsai, C.-T. Lin, G. C. Tseng, I.-F. Chung, and N. R. Pal, "Discovery of dominant and dormant genes from expression data using a novel generalization of SNR for multi-class problems," *BMC Bioinformatics*, Vol. 9, pp.1-33, 2008.

[7] Y.-S. Tsai, K. Aguan, N. R. Pal, and I.-F. Chung, "Identification of single and multiple-class specific signature genes from gene expression profiles by group marker index," *PLoS ONE*, Vol. 6, pp. e24259, 2011.

[8] N.K. Kamila, L.D. Jena, and H.K. Bhuyan, "Pareto-based multi-objective optimization for classification in data mining. *Cluster computing (Springer)*," Vol. 19, No. 4, pp. 1723–1745, Dec 2016.

[9] Jun Wang, Jin-Mao Wei, Zhenglu Yang, and Shu-Qin Wang, "Feature Selection by Maximizing Independent Classification Information" *IEEE Transactions on Knowledge and Data Engineering*, Vol. 29, No. 4, pp. 828 – 841, April, 2017.

[10] H. K. Bhuyan, and N.K. Kamila, "Privacy preserving Sub-feature Selection based on fuzzy probabilities," *Cluster computing, (Springer)*, Vol. 17, No. 4, pp. 1383-1399, 2014.

[11] H. K. Bhuyan, and N.K. Kamila, "Privacy preserving sub-feature selection in distributed data mining," *Applied soft computing, Elsevier*, Vol. 36, pp. 552-569, 2015.

[12] Z. Li, J. Liu, Y. Yang, X. Zhou, and H. Lu, "Clustering-guided sparse structural learning for unsupervised feature selection", *IEEE Trans. Knowl. Data Eng*., Vol. 26, No. 9, pp. 2138-2150, Sep. 2013.

[13] D. Koller, and M. Sahami, "Toward optimal feature selection", *in Proc. 13th Int. Conf. Mach. Learn*., pp. 284-292, 1996.

[14] M. Banerjee, and N. R. Pal, "Feature selection with SVD entropy: Some modification and extension", *Inf. Sci*., Vol. 264, pp. 118-134, 2014.

[15] P. Mitra, C. A. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity", *IEEE Trans. Pattern Anal. Mach. Intell*., Vol. 24, No. 3, pp. 301-312, Mar. 2002.

[16] N. Søndberg-madsen, C. Thomsen, and J. M. Pea, "Unsupervised feature subset selection", *in Proc. Workshop Probabilistic Graph. Models Classification*, pp. 71-82, 2003.

[17] J. Tang, and H. Liu, "Unsupervised feature selection for linked social media data", *in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, pp. 904-912, 2012.

[18] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection", *in Proc. Adv. Neural Inf. Process. Syst*., pp. 507–514, 2005.

[19] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multicluster data", *in Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, pp. 333-342, 2010.

[20] Danilo Costarelli, "Sigmoidal Functions Approximation and Applications," PhD, Dissertation, Dipartimento di Matematica e Fisica Sezione di Matematica, Roma TRE Universita, Deglistudi, 2014.

[21] Monami Banerjee and Nikhil R. Pal, "Unsupervised Feature Selection with Controlled Redundancy (UFeSCoR)," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 27, No. 12, Dec 2015.

[22] H. K. Bhuyan, and C. V. Madhusudan Reddy: Sub-feature selection for novel classification, *IEEE Explore*, April, 2018.