International Journal of Research and Reviews in Applied Sciences And Engineering (IJRRASE) Vol 8. No.1 – 2016 Pp.17-24

available at : www.ijcns.com
ISSN: 2231-0061

FEATURE SELECTION USING MULTIVARIATE ADAPTIVE REGRESSION SPLINES

D.Senthil Kumar
Assistant Professor
CSE Department, Anna University,
BIT-Campus, Trichy.

S.Sukanya
M.E.(Mobile and Pervasive Computing)
Anna University, BIT-Campus, Trichy

ABSTRACT: Multi-label learning is a supervised learning method in which classification algorithm is required to learn a set of instances; where each instance belongs to multiple classes. Feature selection in the multi-label dataset is a challenging task due to complex interaction among features and class labels. Therefore, the multivariate adaptive regression spline (MARS) is used to classify and to select the important features. MARS handles large dataset and makes prediction quickly. Here, an optimistic feature subset selection (OFSS) algorithm has been used. OFSS is applied for selecting the best features from the dataset. The experimental study for this proposed algorithm performs efficiently to predict accurate data.

Keywords: Multi-label classification, Feature selection, Prediction, Multivariate Adaptive Regression Spline.

I. INTRODUCTION:

Nowadays healthcare industry generates massive amount of data about patients. Data analysis is essential for medical decision making and supervision. Analyzing and processing the enormous amounts of data generated by healthcare industry are too complex by conventional method. In data mining, classification is a technique used to predict the target classes accurately for each case. Prediction is a technique used to predict the future from the historical facts [1]. In the healthcare sector, disease caused by a particular symptom cannot be well depicted using a single-label dataset, whereas multi-label dataset can be used to resolve this problem. Multi-label classification is the classification crisis where multiple labels should be assigned to each instance. Multi-label classification was mainly provoked by the task of text categorization, music, and medical analysis. Multi-label classification uses two methods for classification. First, problem transformation method which converts the multi-label problem into a set of binary classification problem then the problem can be handled by the single-label classifier. Second, algorithm adaptation method, adapts the algorithm directly to perform the multi-label classification [2]. In multi-label, the main issue is to select the features for multiple classes. Features available in the multi-label dataset are entirely dependent on all the class labels. The feature selection is the process of selecting the relevant features which are the subset of the features. The features present in the dataset can be used to classify the data with accurate prediction [3], [4]. For this, multivariate adaptive regression splines (MARS) tool is used to handle the complex data and for the selection of optimistic feature subset for the multi-label data.

The rest of this paper is organized as follows. Section II, discuss the related works regarding the

feature selection for multi-label classification and Multivariate adaptive regression spline. In section III, formal definition of multi-label learning is specified. Section IV, introduces the evaluation metrics used in multi-label learning. Section V, it is briefly discusses about the multivariate adaptive regression spline. Section VI, describes the proposed algorithm and result analysis for the multi-label data. In section VII, the main contribution of this paper is summarized.

II. Related Works:

Kwak *et al* [5] proposed two algorithms for feature selection. Greedy selection algorithm and Taguchi algorithm combined together for selecting the features for classification problems. Liu *et al* [6] illustrated with an example how existing algorithm can be integrated into a meta algorithm using the real-world applications. Zhang *et al* [7] presented k-nearest neighbor algorithm for multi-label classification in which test instance of the set of labels are predicted by using maximum a posteriori (MAP) principle.

Lee *et al* [8] proposed credit card scoring using a two-stage hybrid modeling with artificial networks and MARS. MARS build the credit card scoring model and obtain the significant variables then these variables are served as input nodes for the neural network model. Zareipour *et al* [9] proposed that MARS technique can be used for forecasting the hourly ontario energy price (HOPE). The models generated for the HOPE are more accurate and demonstrate the MARS capability electricity marketing price.

Street *et al* [10] proposed a streaming ensemble algorithm for large scale classification that classifies and built a single-label decision tree for all data. For that it requires constant memory and adjusts concept drift. Polikar et al [11] presents that individual classifiers are combined together and the final decision is apparently the most informed one. Boosting, bagging, adaboosting are popular ensemble based algorithm. Thabtah *et al* [12] proposed a new technique called multi-class, multi-label associative classification approach (MMAC). This approach used three measures for evaluating the accuracy for classification. It is highly effective and scalable when compared with other classification approaches.

III. Multi-label Learning

Let S=Z^d denotes instances of the domain and let y = $\{k_1, k_2, \ldots, k_q\}$ be the finite labels. Consider a distribution D for a given training set T = $\{(s_1,Y_1), (s_2,Y_2)\ldots,(s_n,Y_n)\}$ ($s_i \in$ S, $Y_i \subseteq$ y). The function $h : S \rightarrow 2^y$ is used for optimization of the specific evaluation metric. In many cases, multi-label learning system produces a real-valued function of the form $f : S * y \rightarrow$ Z. Let s_i be the instances of the set and Y_i be the associated label, if $f(s_i , y_1) > (s_i , y_2)$ for any $y_1 \in Y_i$ and $y_2 \notin Y_i$ then the labels of Y_i has larger output values.

As stated above, the multi-label problems are inevitable to solve the traditional multi-class problems. Research on multi-label learning was mainly concern the problems of text classification, bioinformatics and scene classification. McCallum *et al* [13] proposed a Bayesian classification approach for multi-label document classification with a mixture of probabilistic model, which generates word distribution for each document. The mixture weights and the word distribution in each mixture component can be determined by using the expectation maximization (EM) algorithm. Matthew R *et al* [14] presented a problem of scene classification where natural scene may contain multiple objects that scene can be described by multiple class labels. In the scene classification, training and testing can be evaluated by precision, recall and overall accuracy. Cheng *et al* [15] proposed probabilistic classifier chain method to look at the problem of risk minimization and Bayes optimal solution for the multi-label classification. The multi-label classification also uses support vector machine (SVM) for yeast gene functional classification, which generates accurate results than others [16].

IV. Evaluation Metrics

The evaluation of multi-label learning methods requires different measures than those used in the case of single-label data. The following evaluation metrics are used for multi-label classification:

(1)*Mean squared error (MSE)*: evaluates the squared units of predicted and observed values, MSE is defined as follows:

$$\text{MSE}= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2 \quad (1)$$

where $\hat{Y}_i$ is a predicted value and Y_i is an observed value.

(2)*Generalized Cross-Validation*: evaluates the best features for the given dataset, GCV is defined as follows:

$$\text{GCV}= \frac{ASR}{(1-M/N)^2} \quad (2)$$

where,

ASR= $(1/N)\sum_{i=1}^{N}\left(y_i - f_M\left(s_i, \hat{\theta}\right)\right)^2$ is the average-squared residual, M is the number of parameters and N is the sample size, $\hat{\theta}$ is an unknown parameters.

(3) *R-squared* (*R^2)*: evaluates how the future samples can be predicted by the model then theR^2 is defined as follows:

$$\text{R}^2=1-\frac{\left(\sum_{i=0}^{n-1}\left(Y_i-\hat{Y}_i\right)^2\right)}{\sum_{i=0}^{n-1}(Y_i-\bar{Y}_i)^2} \quad (3)$$

Where $\hat{Y}_i$ is a predicted value, Y_i is an actual value and

$\bar{Y}_i= \frac{1}{n}\sum_{i=0}^{n-1} Y_i$

(4)*Receiver operating characteristic (ROC)*: A receiver operating characteristic curve is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. A ROC space is defined by FPR and TPR as *x* and *y* axes respectively, which depicts relative trade-offs between true positive rate (TRP) and false positive rate (FRP). Formally, given the number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN), TPR and FPR are calculated as follows:

$$\text{TPR}= \frac{TP}{TP+FP} \quad (4)$$

$$\text{FPR}= \frac{FP}{FP+TN} \quad (5)$$

(5)*Misclassification rate (MCR)*: evaluates the error rate for the predicted values, which is defined as follows:

$$\text{MCR}=1-\frac{\sum_{i=1}^{n} D_{ii}}{\sum_{i=1}^{n} D_{ij}} \quad (6)$$

Where, D_{ii} is the sum of diagonal elements of the confusion matrix, D_{ij} is the sum of all elements in the confusion matrix.

V. MARS

MARS is a modeling tool mainly for statistical analysis and in data mining. MARS will randomly partition 80% of the data as training samples and 20% of the data as testing samples. The multi-label data contains attributes with multiple classes which are converted into single-label classes. Each class of the target variable and its attributes are given as input to MARS then it creates the model and graphically displays collision of each predictive aspect on the outcome. MARS builds the model in the form:

$$\widehat{f}(x)=\sum_{m=1}^{k} c_m B_m(x) \quad (7)$$

where, k is a number of variables, $B_m(x)$ is a basic function and c_m is a constant coefficient [17]. Basic function has any one of the three forms: a) a constant 1or b) hinge function or c) product of more than two hinge function. Hinge function may be $max(0, x-c)$ or $max(0, c-x)$. The modeling process has two stages. The forward pass finds pair of basic functions at each step, reduces the maximum sum-of-squares residual error. It builds an over-fit model and the backward pass builds the model with better generalization capability, at each step it prunes the model with least effective terms until it finds the best sub model. Backward pass uses GCV to compare the performance of model subsets in order to choose the best subset features. After selecting the features combine all the

features of each class and sort the features according to the weight of the variable. The subsets of features are predicted for the selected features using the threshold function. Consider the default threshold value as .5 for all the target values.

If the threshold value less than .5 corresponds to non-response value and it is greater than .5 corresponds to response values.

VI. Proposed Algorithm:

The proposed algorithm selects the subset feature for the multi-label dataset. It calculates the generalized cross validation for selecting the best features and by sorting the variables using its weight. Then evaluate the features using the threshold value. An optimized feature subset selection algorithm is shown below:

Algorithm (optimistic feature subset selection)
Input : Multi-label Dataset (MLD)
Output : Selected features of subset

Step1: Dividing the multi-label data into single-label data.
Step2: Select the features by using mars for each class. In order to choose the best features using generalized cross validation method.

$$GCV=\sum_{i=1}^{N}\frac{(y_i-f(x_i))^2}{(1-{}^{c}/_{N})^2} \qquad (8)$$

C=1+cd
where, N is the number of cases in the dataset.
d effective degree of freedom.
c is the penalty for adding a basic function.
y_iVariable to be predict.
$f(x_i)$ Predicted value of y_i .

Step3: Selected features of multi-label dataset:
1) By using, $\bigcup_{i=1}^{n} c_i(Fs)$ unite all the selected features of each class.

Where, c_i = number of class in the dataset.
2) Sort the features according to the weight of the variable.

Step4: Select top s= $\log_2 n$
where, s=number of features using the threshold function.
n=number of features in the overall multi-label dataset.

A. Result and Discussion:

This section illustrates the accuracy and evaluation metrics for the selected features of the multi-label dataset. The coronary heart disease dataset contains 181 features, 555 objects and six classes.

Table. 1. Accuracy for the coronary heart disease

Pattern of syndrome	MARS with raw dataset	Optimized dataset
Deficiency of heart Qi	81.10%	67.87%
Deficiency of heart Yang	66.48%	64.66%
Deficiency of heart Yin	63.64%	58.29%
Qi stagnation syndrome	42.20%	36.59%
Phlegm	42.20%	57.14%
Blood stasis	84.20%	88.64%

The proposed algorithm selects 7 best features for the optimized dataset using MARS tool. Table 1 shows the accuracy for the coronary heart disease between the raw dataset and the optimized dataset for each class using multivariate adaptive regression spline tool.

Table 2 shows that evaluation metrics for each class of raw dataset. Table 3 infers the evaluation metrics for each class of optimized dataset and Fig 1 shows the accuracy for the coronary heart disease of raw and optimized dataset. Fig 2 represents the evaluation metrics for the class deficiency of heart Qi. Fig 3 represents the evaluation metrics for the class deficiency of heart yang. Fig 4 represents the evaluation metrics for the class deficiency of heart Yin. Fig 5 represents the evaluation metrics for the class deficiency of Qi stagnation syndrome. Fig 6 represents the evaluation metrics for the class Phlegm. Fig 7 represents the evaluation metrics for the class Blood stasis.

Table. 2. Evaluation metrics for raw dataset using MARS

Pattern of syndrome	MSE	GCV	R^2	ROC	MCR
Deficiency of heart Qi	0.183	0.207	0.228	0.773	0.288
Deficiency of heart Yang	0.139	0.172	0.353	0.837	0.190
Deficiency of heart Yin	0.197	0.216	0.167	0.723	0.295
Qi stagnation syndrome	0.133	0.148	0.188	0.743	0.171
Phlegm	0.210	0.230	0.157	0.716	0.344
Blood stasis	0.163	0.171	0.103	0.691	0.239

Table 1 and 2 shows the experimental result of five evaluation metrics for the classes of coronary heart disease for raw dataset and optimized dataset. The dataset contains six classes which are denoted as patter of syndrome in the table. The evaluation metrics are calculated for each class of the raw dataset and the optimized dataset. MSE is the missing values between the simulated values and the observed values of each class of the raw and optimized dataset.

Pattern of syndrome	MSE	GCV	R^2	ROC	MCR
Deficiency of heart Qi	0.218	0.222	0.082	0.617	0.322
Deficiency of heart Yang	0.185	0.191	0.139	0.664	0.255
Deficiency of heart Yin	0.214	0.224	0.095	0.644	0.329
Qi stagnation syndrome	0.158	0.163	0.036	0.573	0.200
Phlegm	0.240	0.244	0.037	0.596	0.405
Blood stasis	0.177	0.180	0.027	0.582	0.239

Table. 3. Evaluation metrics for optimized dataset using MARS

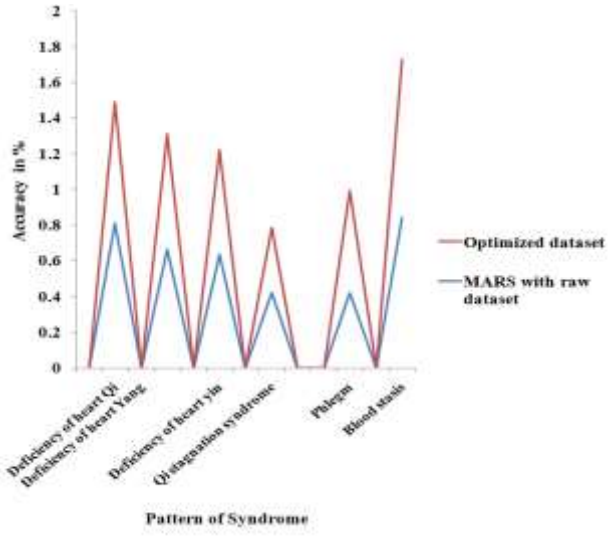

Fig.1 shows accuracy for the coronary heart disease

The MSE value of the optimized dataset has highest value 0.240 before that the MSE value for raw dataset is 0.210.

GCV is calculated for choosing the best features of each class in the dataset. The Generalized Cross Validation of the optimized dataset is 0.244 and the raw dataset value is 0.230. R^2 is the fraction by which the variance of errors is less than the variance of the dependent variable calculated for each class of the raw dataset and the optimized dataset. R-squared has the highest value 0.353 for the raw dataset whereas the optimized dataset has the R-squared value as 0.139. It is difficult to calculate variance of the error. For multiple models, it determines the correlation among all the variables such as dependent and independent variables. ROC curve is plotting the true positive rate against the false positive rate for each class value of the raw dataset and optimized dataset. The ROC has the highest value 0.837 for raw dataset and 0.664 for optimized dataset. MCR calculates the error rate of misclassified features for the each class in the raw dataset and the optimized dataset. Misclassification error rate has the highest value 0.405 for the optimized dataset and the raw dataset has the value as 0.295. The value of each evaluation metrics

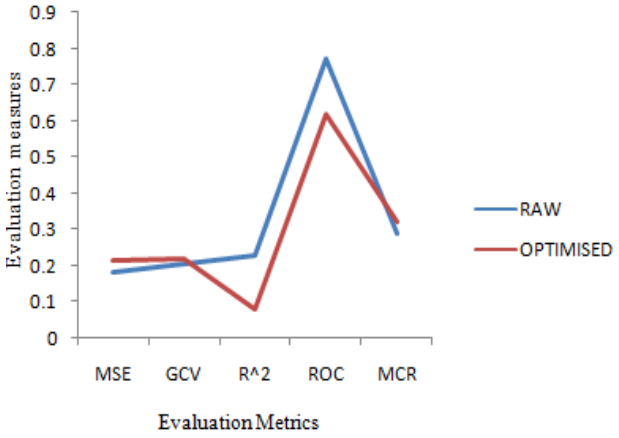

Fig. 2. Evaluation metrics of raw and optimized data for the class deficiency of heart Qi

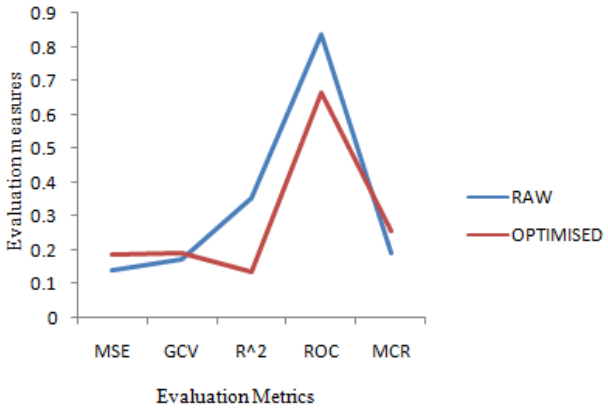

Fig. 3. Evaluation metrics of raw and optimized data for the class deficiency of heart Yang

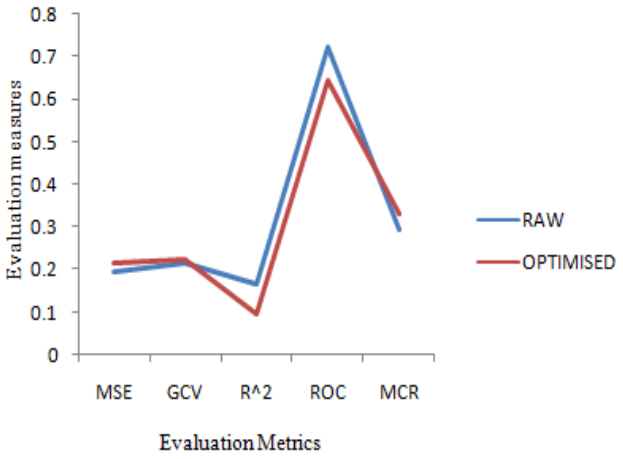

Fig. 4. Evaluation metrics of raw and optimized data for the class deficiency of heart Yin

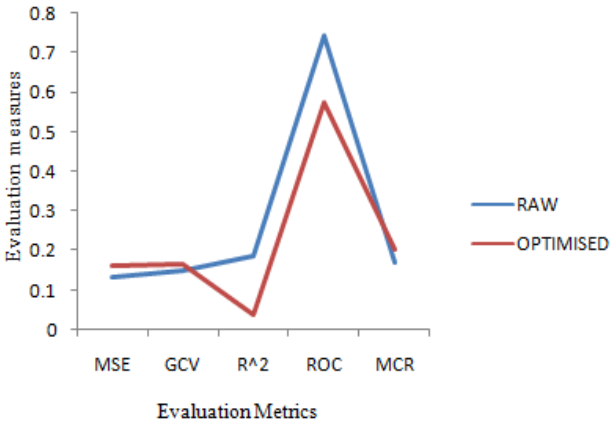

Fig. 5. Evaluation metrics of raw and optimized data for the class Qi stagnation syndrome

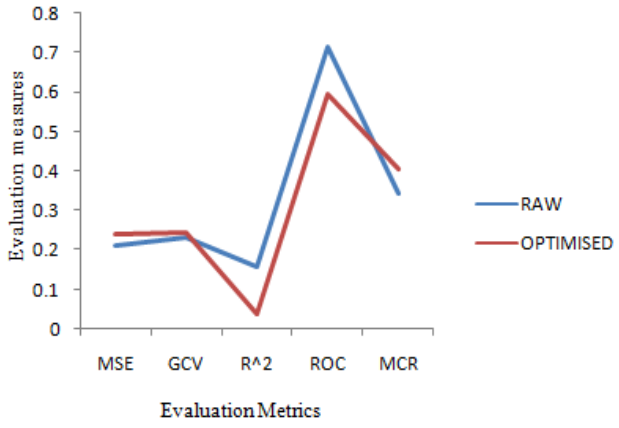

Fig. 6. Evaluation metrics of raw and optimized data for the class Phlegm

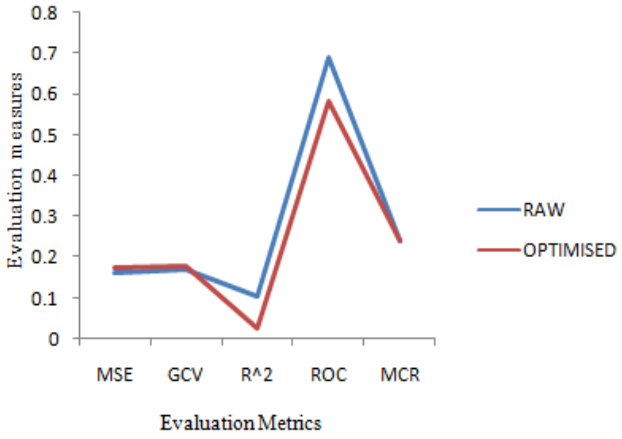

Fig. 7. Evaluation metrics of raw and optimized data for the class Blood stasis

for every class has the probability of error rate between0 to 1. On comparing raw and optimized dataset, raw dataset produces better results.

VII. CONCLUSION:

This paper presents the optimized feature subset selection for the multi-label dataset using MARS. It is tedious to select the optimized features for multi-label dataset because it contains more number of features with multiple classes. Here, the MARS tool can be used to predict the large dataset. The proposed algorithm has been used to select the best features for the multi-label dataset. As a result, the experimental study produces better accuracy and evaluation metrics for the raw dataset than the optimized dataset but still accuracy and some of the evaluation metrics for the classes in the dataset need to be improved.

REFERENCES

[1] Han, Jiawei, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques*. Elsevier, 2011.

[2] Tsoumakas, Grigorios, and Ioannis Katakis. "Multi-label classification: An overview." *Dept. of Informatics, Aristotle University of Thessaloniki, Greece* (2006).

[3] Dash, Manoranjan, and Huan Liu. "Feature selection for classification." *Intelligent data analysis* 1.3 (1997): 131-156.

[4] Zhang, Min-Ling, José M. Peña, and Victor Robles. "Feature selection for multi-label naive Bayes classification." *Information Sciences* 179.19 (2009): 3218-3229.

[5] Kwak, Nojun, and Chong-Ho Choi. "Input feature selection for classification problems." *Neural Networks, IEEE Transactions on* 13.1 (2002): 143-159.

[6] Liu, Huan, and Lei Yu. "Toward integrating feature selection algorithms for classification and clustering." *Knowledge and Data Engineering, IEEE Transactions on* 17.4 (2005): 491-502.

[7] Zhang, Min-Ling, and Zhi-Hua Zhou. "A k-nearest neighbor based algorithm for multi-label classification." *Granular Computing, 2005 IEEE International Conference on*. Vol. 2. IEEE, 2005.

[8] Lee, Tian-Shyug, and I-Fei Chen. "A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines." *Expert Systems with Applications* 28.4 (2005): 743-752.

[9] Zareipour, H., K. Bhattacharya, and C. A. Canizares. "Forecasting the hourly Ontario energy price by multivariate adaptive regression splines." *Power Engineering Society General Meeting, 2006. IEEE*. IEEE, 2006.

[10] Street, W. Nick, and YongSeog Kim. "A streaming ensemble algorithm (SEA) for large-scale classification." *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001.

[11] Polikar, Robi. "Ensemble based systems in decision making." *Circuits and systems magazine, IEEE* 6.3 (2006): 21-45.

[12] Thabtah, Fadi A., Peter Cowling, and Yonghong Peng. "MMAC: A new multi-class, multi-label associative classification approach." *Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on*. IEEE, 2004.

[13] McCallum, Andrew. "Multi-label text classification with a mixture model trained by EM." *AAAI'99 workshop on text learning*. 1999.

[14] Boutell, Matthew R., et al. "Learning multi-label scene classification." *Pattern recognition* 37.9 (2004): 1757-1771.

[15] Cheng, Weiwei, Eyke Hüllermeier, and Krzysztof J. Dembczynski. "Bayes optimal multilabel classification via probabilistic classifier chains." *Proceedings of the 27th international conference on machine learning (ICML-10)*. 2010.

[16] Elisseeff, André, and Jason Weston. "A kernel method for multi-labelled classification." *Advances in neural information processing systems*. 2001.

[17] Barron, Andrew R., and Xiangyu Xiao. "Discussion: Multivariate Adaptive Regression Splines". *The Annals of Statistics* 19.1 (1991): 67–82.