Intelligent classification model for railway signal equipment fault based on SMOTE and ensemble learning

Lianbao YANG*, Ping LI, Rui XUE, Xiaoning Ma, Xinqin LI and Zhe WANG

China Academy of Railway Sciences, Beijing, 100081, China

*E-mail: yanglianbao_121@163.com

Abstract. In this paper, we propose a novel intelligent classification model to classify the railway signal equipment fault based on SMOTE and ensemble learning. To tackle the imbalanced fault text data, the model uses SMOTE algorithm to generate the minority railway signal equipment fault class data randomly, making the data balanced. Then the model adopts the base classifier, such as Logistic Regression, Multinomial Naive Bayes, SVM and the ensemble classifier, such as GBDT, Random Forests to classify the data processed by SMOTE. To combine the advantages of various classifiers, the model integrates multiple classifiers by way of voting. Based on the experiment analysis of railway signal equipment fault text data from 2012 to 2016, the result shows that the model has a significant improvement in fault classification accuracy, recall rate and f-score.

1. Introduction

For the various mechanism of different kinds of railway signal equipment, the frequency of fault occurrence is not the same, which causes the problem of imbalanced fault classification. It means that only a small amount of data belong to a kind of railway signal equipment fault, and the vast majority of the data belong another kind of railway signal equipment fault. To avoid the imbalanced fault classification problem, this paper proposes a novel railway signal equipment fault intelligent classification model, which mainly includes two steps. Firstly, the data balance, which means representing the unstructured text data in vector and balance the transformed vector data. Secondly, the EL (Ensemble Learning) [1], which means training and integrating classifiers by the processed data to accomplish intelligent classification of railway signal equipment fault.

The data balance is mainly to balance the data by changing the sample distribution of the data set, which is mainly divided into over-sampling and under-sampling [2]. Over-sampling is to generate small class data automatically, and under-sampling is to remove data from the large category. In 2002, Chawla proposed the SMOTE (Synthetic Minority Oversampling Technique) [3] algorithm which is commonly used in the over-sampling methods. The SMOTE algorithm includes several improved versions, such as the Borderline-SMOTE [4], SVM-SMOTE. The basic idea of SMOTE is to synthesize some new data in minority classes to achieve the balance. The ensemble learning is mainly to make full use of the difference of classifiers through training multiple classifiers and implement the integration of different classifiers through Voting. Traditional classification is primarily based on a single classifier model, such as LR (Logistic Regression), DT (Decision Tree), SVM (Support Vector Machine) and Multinomial NB (Multinomial Naive Bayesian) [5]. However, these classifier models are mainly suitable for training on balanced data. Ensemble classifiers mainly consist of bagging and boosting methods. RF (Random Forest) [6] is the main representative algorithm of bagging method and GBDT (Gradient Boost Decision Tree) [7] is the representative algorithm of the boosting method.

In view of experts and scholars research of imbalanced data classification, and combined with the characteristics of railway signal equipment fault data, this paper puts forward a novel multiple classifiers ensemble learning railway signal equipment fault intelligent classification model based on the SMOTE and Voting. The model utilizes SMOTE algorithm to implement small categories of railway signal equipment fault data automatically, and integrates base classifiers such as LR, Multinomial NB, SVM and ensemble classifiers such as RF, GBDT by means of Voting to achieve railway signal equipment intelligent classification. In order to validate the correctness and effectiveness of the model, a total of 10 classed 641 records for the railway signal equipment fault in the railway station from 2014 to 2016 are analyzed.

2. Imbalanced fault text data processing based on SMOTE

SMOTE is a kind of over-sampling technique which is used to synthesize the minority class data to meet the balance of the data set, which can enhance the effect of classification. Its basic principle is: by selecting a minority sample x_i and k nearby similar samples, select x_j from nearby randomly of its class, and by random linear interpolation, constructs the new minority sample x_{new}:

$$x_{new} = x_i + u\left(x_i - x_j\right), 0 \le u \le 1 \tag{1}$$

Since traditional SMOTE does not consider the distribution characteristics of its adjacent samples, it may cause that repetition occurs between classes. In recent years, some improved algorithms based on SMOTE have been proposed, and the representative algorithms include the Borderline-SMOTE algorithm and SVM-SMOTE algorithm, etc. Borderline-SMOTE only makes linear interpolation on a few samples of the boundary, thus enhancing the impact of boundary samples. SVM - SMOTE is based on the proximity ratio of different types of samples, and the classification boundary can be constructed through SVM, which can be interpolated according to the actual sample data distribution, making the distinction between categories more obvious. Therefore, this paper selects SVM - SMOTE algorithm to generate data in minority classes. The effect of different SMOTE algorithms on the generation of samples in minority is shown in figure 1. The blue dots in figure 1 represent the minority samples, and the red dots represent the majority samples.

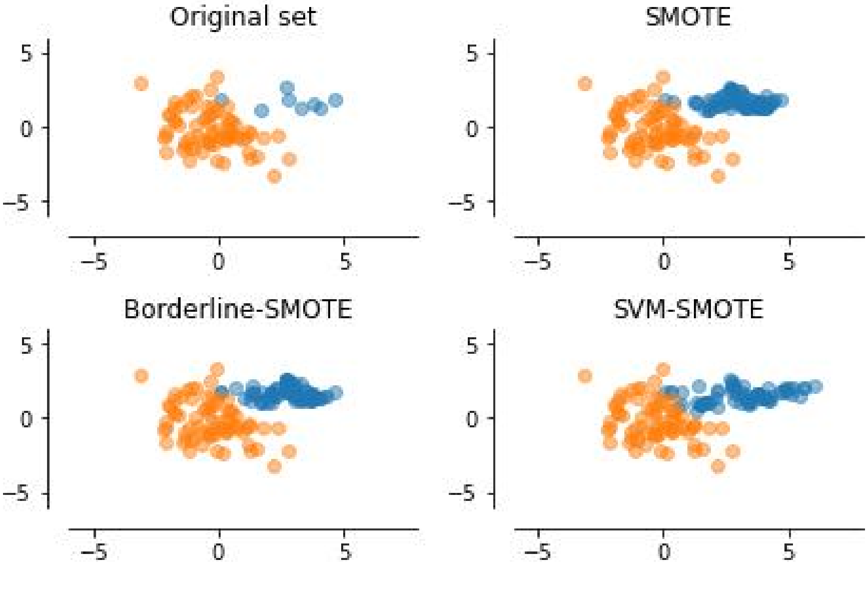

Figure 1. Different SMOTE algorithms used to synthesize the minority samples.

3. Voting based intelligent classification model of multi-classifier ensemble learning

3.1. Base classifier

The commonly used text classifiers are LR, DT, SVM and Multinomial NB, etc. Suppose n vector data set with a m dimension feature, which has c classification, can be expressed as: $D=\{x_i,y_i\},i=1,2,...,n\,,x_i=(x_i^1,x_i^2,...,x_i^m)\,,y_i\in\{0,1,2,..c-1\}$

3.1.1. LR.

LR is a kind of classification method based on statistical analysis.

$$p(y=k\mid x)=\frac{e^{-g_k(x)}}{1+\sum_{j=0}^{c-1}e^{-g_j(x)}},k=0,1,2,\cdots,c-1 \tag{2}$$

Therefore, the corresponding Logistic regression model can be obtained:

$$g_k(x)=\beta_{k0}+\beta_{k1}x^1+\beta_{k2}x^2+\cdots+\beta_{km}x^m \tag{3}$$

The calculation of parameters β is usually estimated by the maximum likelihood method

3.1.2. DT.

DT is a special tree structure, mainly used for classification and decision making. A decision tree contains three types of nodes: decision nodes, usually represented by rectangular boxes; opportunity nodes, usually represented by a circle; endpoints, usually represented by triangles. The commonly used decision tree generation algorithms are ID3, C4.5 and C5.0.

3.1.3. SVM.

SVM is constructed by constructing a hyperplane $f(x)$ that allows the function to represent the relationship between the class y and the sample x vector. Definition of linear x insensitive loss function is:

$$\left|y-f(x)\right|_\varepsilon=\begin{cases}0, & \left|y-f(x)\right|\le\varepsilon\\ \left|y-f(x)\right|-\varepsilon, & \left|y-f(x)\right|>\varepsilon\end{cases} \tag{4}$$

If there is a hyperplane:

$$f(x)=\omega^Tx+b=0\;,\omega\in R^n,b\in R \tag{5}$$

Let:

$$\left|y-f(x)\right|\le\varepsilon \tag{6}$$

The sample set D is called a ε -linear, $f(x)$ is regression estimation function. The distance from the sample point $\{x_i,y_i\}$ to the hyperplane is:

$$d_i=\frac{\left|\omega^Tx_i+b-y_i\right|}{\sqrt{1+\|\omega\|^2}}\le\frac{\varepsilon}{\sqrt{1+\|\omega\|^2}} \tag{7}$$

To get the optimal hyperplane classification, it can transform to an optimization problem, as let the $\|\omega\|^2$ to be minimum. But for nonlinear problems, SVM maps samples to a high-dimensional feature space through nonlinear mapping function $\varphi(x)$, and computes inner product through kernel function. The optimization problem can be expressed as:

$$\min(\frac{1}{2}\|\omega\|^2+\frac{C}{n}\sum_{i=1}^{n}(\xi_i+\xi_i^*))$$
$$s.t\begin{cases}\omega^T\varphi(x_i)+b-y_i\le\varepsilon+\xi_i\\ y_i-\omega^T\varphi(x_i)-b\le\varepsilon+\xi_i^*\\ \xi_i,\xi_i^*\ge 0\end{cases}\tag{8}$$

In (8), ξ_i , ξ_i^* are the slack variable; C is the penalty factor. The larger penalty factor is, the larger the penalty for sample classification errors, and adjusting C can improve SVM's generalization ability.

3.1.4. Multinomial NB.

Multinomial NB is naive bayes model which is adapted to discrete characteristics. It takes documents as a collection of frequency words, and when calculating the prior probability and conditional probability, smooth will be processed so as to solve posteriori probability is 0 if a characteristic didn't appear in the training sample.

The calculation of prior probabilities $p\left(y=k\right)$ is as follows:

$$p\left(y=k\right)=\frac{N_{y=k}+\alpha}{N+c\alpha}\tag{9}$$

The condition probabilities $p\left(x_i\middle|y=k\right)$ is as follows:

$$p\left(x_i\middle|y=k\right)=\frac{N_{y=k},x_i+\alpha}{N_{y=k}+m\alpha}\tag{10}$$

In (9), (10), $N_{y=k}$ is the number of samples of the k category, $N_{y=k},x_i$ is the number of samples in category k with the eigenvector is x_i . α is the smoothness value, and when $\alpha=1$ it's called the Laplace smooth; when $0<\alpha<1$, it's called Lidstone smooth, when $\alpha=0$, it's means don't do smooth.

3.2. Ensemble classifier

An ensemble classifier is a classifier that makes common decisions by combining multiple base classifiers with a certain strategy. It mainly includes the boosting algorithm which supposes the interdependence of base classifiers, and the bagging algorithm which supposes the independence of each base classifier. Boosting algorithm is to operate the sample set to get the sample subset, then take advantage of sample subset to train base classifiers, and finally get the ensemble classifier through the weighted fusion of trained base classifiers. Bagging algorithm begins with randomly selecting training data to train base classifier, and then combines the different trained base classifiers to get the ensemble classifier. This article selects a parallel ensemble classifier RF, which is based on bagging and a serial ensemble classifier GBDT, which is based on Boosting for text classification.

The RF uses the CART decision tree as the base classifier, and improves the establishment process of decision tree. Decision tree selects an optimal feature make the left and right subtrees partition on all nodes in a sample of n features, but RF randomly picks n_{sub} ($n_{sub}\le n$) features of a sample characteristics of nodes, then chooses an optimal feature to make the left and right subtrees partition, further enhancing the generalization ability of the model and avoiding over-fitting phenomenon. The main tuning parameters of RF are divided into bagging frame parameters and CART decision tree parameters. The bagging frame parameters include the maximum number of iterations, etc. The CART decision tree parameters include the maximum depth of the tree.

GBDT, also called MART (Multiple Additive Regression Tree), is a kind of iteration of the decision tree algorithm. The algorithm is composed of many CART Regression decision trees, through the gradient promotion algorithm optimizing the loss function, and finally gets the optimal regression tree. GBDT tuning parameters are much more, and can mainly divided into boosting framework and CART tree parameters. And the boosting framework parameters include the maximum number of iterations, learning step length, etc. The CART decision tree parameters include the depth of the tree.

3.3. Voting based multi-classifier integration learning

The basic idea of ensemble learning is to build one strong classifier based on some weak classifiers through different strategies, solving classification problem. The advantage of various classifiers ensemble learning is that it solved some statistical problems in reality when using weak classifiers. Through combining results of weak classifiers, ensemble learning can reduce location risk in classification. Therefore, ensemble learning always has a better performance in generalization.

Supposing error rates of classifiers are independent, as the number of base classifiers increases, error rate of ensemble learning decreases exponentially, and finally trends to zero. However, base classifiers are always somehow dependent with each other. To choose relatively accurate and diverse classifiers for ensemble learning, this paper picks classifiers based on their classification performance in sample datasets. And then through a voting system of the best classifier combined with other classifiers, this paper chooses the combination with the best classification performance as a final ensemble classifier. The workflow of this method is shown in figure 2.

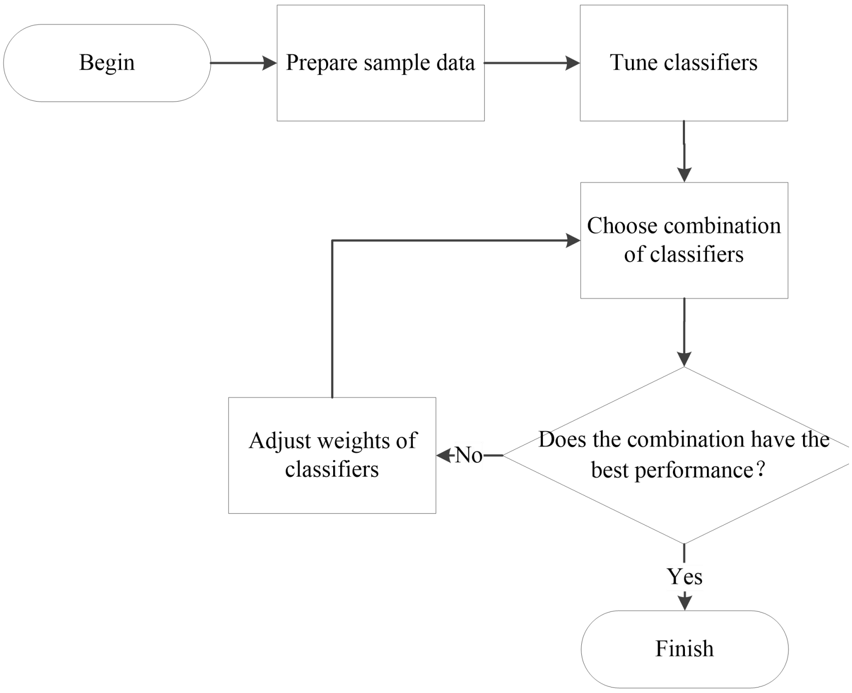
Begin
Prepare sample data
Tune classifiers
Choose combination of classifiers
Does the combination have the best performance?
No
Adjust weights of classifiers
Yes
Finish

Figure 2. Workflow of Various Classifiers Ensemble Learning Based on Voting.

4. Experiment analysis

This experiment uses imbalanced fault data of signal equipment from one railway bureau signal division, to validate the model proposed above. Experiment data include 643 items covering 10 types of faults. Precision, recall and F-score are used as evaluation index of the model.

Precision is defined as:

$$Precision = \frac{1}{|C|}\sum_{i\in C}\frac{(TP_i + TN_i)\times TP_i}{TP_i + FP_i} \tag{11}$$

Recall is defined as:

$$Recall = \frac{1}{|C|}\sum_{i\in C}\frac{(TP_i + TN_i)\times TP_i}{TP_i + FN_i} \tag{12}$$

F-score is formulated as:

$$F_1 = \frac{2\times Precision \times Recall}{Precision + Recall} \tag{13}$$

TP_i (true positive) means the number of items distributed to this class correctly, TN_i (true negative) means the number of items distributed not to this class correctly, FP_i (false positive) means the number of items wrongly distributed to this class, FP_i (false negative) means the number of items belonged to this class while distributed to other classes, and C represents the number of classes.

The experiment includes two parts. The first part is imbalanced fault text data experiment analysis, and the second part is fault intelligent classification experiment analysis.

4.1. Imbalanced fault text data experiment analysis

This experiment uses Jieba to accomplish word segmentation of signal equipment fault text, and calculates weights and normalizes through TF-IDF, and finally represents fault text as vectors. To validate effects of SMOTE when dealing with imbalanced fault text data, this experiment generates some data in minority through SVM-SMOTE. TDCS equipment fault data are generated from origin 6 items to 172 items, blocking equipment fault data are from 6 items to 84 items, and microcomputer interlocking fault data are from 6 items to 133 items. And the overall data become 1014 items, as shown in table 1.

Table 1. Comparison between raw data and data after SMOTE.

Category	Raw Data	Data after SMOTE
CTC equipment fault	21	21
LKJ equipment fault	15	15
TDCS equipment fault	6	172
Blocking equipment fault	6	84
On-board equipment fault	23	23
Turnover fault	175	175
Power supply panel fault	53	53
Track circuit fault	242	242
Computer interlocking fault	6	133
Signal fault	96	96
Total	643	1014

4.2. Fault intelligent classification experiment analysis

This experiment uses two different datasets (raw data and data after SMOTE) to train and test widely used traditional classifiers (LR, Multinomial NB and SVM), and ensemble classifiers (RF and GBDT). And then compares their performance on precision, recall and F-score, to validate SMOTE's effect of imbalanced fault text data classification. From both two datasets randomly take 80% as training data, and 20% as testing data.

4.2.1. Raw text data classification experiment.

Ensemble classifiers need tuning based on training data to achieve a better performance. This experiment tunes through GridSearchCV. After tuning, the estimator of RF is 180, the estimator of GBDT is 180 as well, learning rate is 0.3, and subsample is 0.8. The performance of tuned ensemble classifiers and base classifiers shows in table 2.

Table 2. Raw data classification performance.

	Algorithm	**Precision**	**Recall**	**F1**
Base Classifiers	LR	0.8001	0.7752	0.7357
	Multinomial NB	0.9136	0.8527	0.8745
	SVM	0.8997	0.8527	0.8489
Ensemble Classifiers	RF	0.9078	0.8837	0.8778
	GBDT	0.9036	0.8760	0.8671

Table 3 shows that for imbalanced data, ensemble classifiers are slightly better than base classifiers. And among base classifiers, logic regression has the worst performance in this experiment, while among ensemble classifiers, RF shows the best performance in classification.

4.2.2. Text data classification experiment after SMOTE.

After using SMOTE to balance data, this experiment retrains and retests classifiers. And then the tuned parameters are the following: RF estimator is equal to 160, while GBDT estimator is 180, learning rate is 0.3, and subsample is 0.6. The classification performance after SMOTE is shown in table 3.

Table 3. Data after SMOTE classification performance.

	Algorithm	**Precision**	**Recall**	**F1**
Base Classifiers	SMOTE + LR	0.8963	0.9015	0.8768
	SMOTE + Multinomial NB	0.9312	0.9261	0.9267
	SMOTE + SVM	0.9538	0.9458	0.9416
Ensemble Classifiers	SMOTE + RF	0.9441	0.936	0.932
	SMOTE + GBDT	0.9253	0.931	0.9294

Table 4 describes that after SMOTE, the performance for both base classifiers and ensemble classifiers is significantly improved, especially SVM.

4.2.3. SMOTE + Voting ensemble classification experiment.

Based on above experiments, choose some classifiers to ensemble based on voting. Among those combinations, ensemble learning based on SVM with other four classifiers voting stands out. And the performance shows in table 4.

Table 4. SMOTE + Voting ensemble classification performance.

	Algorithm	Precision	Recall	F1
SMOTE + Voting Ensemble Learning	SVM + LR + RF	0.9538	0.9458	0.9416
	SVM + LR + GBDT	0.9538	0.9458	0.9416
	SVM + LR + Multinomial NB	0.9538	0.9458	0.9416
	SVM + Multinomial NB+ RF	0.9607	0.9557	0.9528
	SVM + Multinomial NB + GBDT	0.9572	0.9507	0.9474
	SVM + GBDT + RF	0.9517	0.9458	0.9425

Table 4 demonstrates that combinations such as SVM + LR + RF, SVM + LR + GBDT, SVM + LR +Multinomial NB do not improve the final performance, which is just like SVM alone. While ensemble learning with various classifiers such as SVM + Multinomial NB + RF, SVM + Multinomial NB + GBDT, SVM + GBDT + RF has a better performance compared to using SVM alone. And SVM + Multinomial NB + RF performs the best, which improves 1% on precision, recall and F-score.

4.3. Experiment conclusion

Based on above experiment analysis, take SVM classifier as an example, and compare performance of using raw data, data after SMOTE and data after SMOTE + Voting, shown in figure 3.

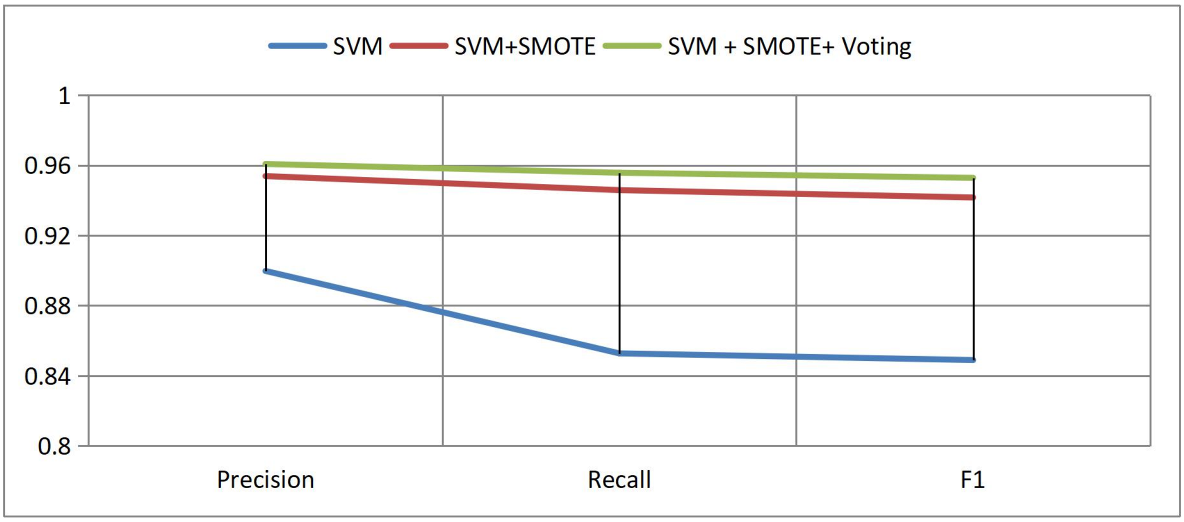

Figure 3. SVM performance comparison of different methods.

From figure 3 it shows that for imbalanced data, SVM+ SMOTE+ Voting method shows the best performance in precision, recall and F1. Therefore, intelligent classification model for railway signal equipment fault text data, which is firstly proposed in this article, has great advantages and can be applied in railway signal equipment.

5. Conclusion

This article uses SVM-SMOTE to automatically generate data in minority at first, balancing fault text data and improving classification performance. And then it applies these relatively balanced new text data to train and test models, and chooses models with competitive performance. Models used as classifiers include base ones such as LR, Multinomial NB, SVM, and ensemble ones such as RF and GBDT. Finally based on those models, this paper proposes a new ensemble learning classification model based on voting method. This article also includes an experiment dealing with imbalanced fault text data from railway signal equipment of one railway bureau. The experiment validates the model mentioned above through precision, recall and F-score, offering new solution to intelligent classification of railway signal equipment faults.

Acknowledgement

The authors would like to thank the reviewer's great patient to revise the paper and the support of the China Academy of Railway Sciences within the major issues of the fund (no. 2017YJ005 and no.2017YJ02).

References

[1] Dietterich T G 2000 *Mutliple classifier systems*, pp:1-15
[2] He H, Bai Y, Garcia E A, Li S 2008 *IEEE International Joint Conference on Neural Networks*. pp:1322-28.
[3] Chawla N V, Bowyer K W, Hall L O, Kegelmeyer W P 2002 *Journal of Artificial Intelligence Research*, 16: 321-57.
[4] Han H, Wang W Y, Mao B H 2005 *Lecture Notes in Computer Science* 3644:878-887.
[5] Aggarwal C C, Zhai C X 2012 *Mining Text Data.*pp:163-222.
[6] Breiman L 2001 *Machine Learning* 45 (1): 5-32.
[7] Friedman J H 2001 *Annals of Statistics* 29 (5):1189-232.