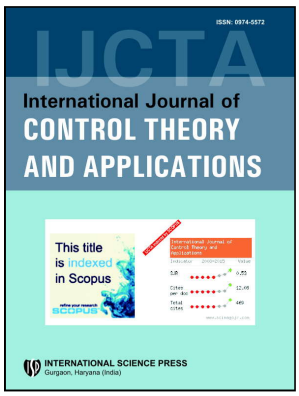

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 14 • 2017

A Novel Cluster of Feature Selection Method Based on Information Gain

Sai Prasad Potharaju[1] and M. Sreedevi[2]

[1] *Dept of Computer Science and Engineering K L University, Guntur (AP), India, Email: psaiprasadcse@gmail.com*
[2] *Dept of Computer Science and Engineering K L University, Guntur (AP), India, Email: msreedevi_27@kluniversity.in*

Abstract: Data Mining is one of the fastest growing data sciences in the information technology world. Classification is one of the functions of data mining. In addition to classification, the functions of Data mining can be divided into feature selection, clustering and association rule mining. Feature selection is one of the important functions among these, and it is a preprocessing technique. The classification algorithm accuracy is calculated by how correctly a class can be predicted on a given dataset. The accuracy of an algorithm depends on choosing of correct attributes or features from a given dataset. Feature selection is important to record the better performance of an algorithm. Information gain is one of the techniques to choose the features from the dataset. Feature selection methods used for getting accurate features from the given dataset in order to reduce the computation time. In this paper, we present a novel feature selection method using information gain. In our method, defined number of features of original dataset are divided into a number of clusters (sets) based on the rank or weight of an attribute .Each cluster contains set of features and these clusters are analyzed on different classification algorithm with different feature selection methods. The result of each cluster of features is recorded and compared with existing feature selection methods. The result of one of the clusters is competing with others Feature selection method.

Keywords: Classification, Data Mining, Feature Selection, Information Gain, SMOTE

1. INTRODUCTION

In recent days all the organizations are taking the advantages of data mining. The production of data has been increasing with high speed rate by the various divisions of organizations. In such cases, data mining functions like pre-processing, Assessment, Feature selection, Clustering, Classification, Association rule mining plays an important role. A significant number of parallel and distributed techniques have been proposed by researchers to mine frequent item sets based on a support threshold of item set [1].Data mining methods can be considered based on different disciplines that include computer science, operational research, statistics and machine learning [2].

Classification is widely accepted tool to handle the many real time problems, which includes fraud detection, virus detection, disease detection, student performance detection, crime detection and many. Classification

technique decides, a data belongs to which class by analyzing a training dataset. Before undergoing training phase, the dataset should be preprocessed. In this phase, noise will be removed if present, missing values will be normalized. Feature selection is one of the pre-processing techniques. The best features will give better result of the classification algorithm. There are various feature selection approaches and algorithms available in recent studies, which will be discussed in subsequent sections.

In literature, there are various classification methods such as SVM-support vector machine, C.5.0- Decision tree, rough set, NN-neural network, etc. [3]. A method for classification of multiple classes using binary learning techniques based on margin is discussed in the article [4]. In the research [5] , a new method of NPVM- nonparallel support vector machine with universum learning for classification of samples that are not belonging to any class was designed.

Authors of [6], discussed multiclass text classification using Naive Bayes (NB) and SVM. Different approaches are used to combine the binary classifiers, the comparison results shows that SVMs performance is better than Naive Bayes. In [7], researchers proposed CFNN (cascade-forward NN) approach for classifying electricity price. In this article [7], researchers proposed CFNN method which provides a robust and accurate method for electricity market price classification classes. The authors of [8], worked on prediction of kidney disease using different rule based and tree based algorithms using SMOTE [9].Classification result without using SMOTE and using SMOTE is compared in their article. The result was increased using SMOTE which is over sampling technique.

However, the objective of this current study is to describe a novel feature selection approach based on information gain. Our proposed method has been analyzed by three datasets, among those two datasets are balanced using SMOTE, which is a sampling technique, and another one is imbalanced dataset.

The remaining section of this article is organized as follows. In section II, we discuss about related work and basic methods of feature selection. In section III, we present our proposed method with a simple example for better understanding. In section IV, we present experiments on datasets with its results. In Section V, we provide conclusion with future work

2. RELATED STUDY

The feature selection is the process of choosing subset of features or attributes from the given dataset which can influence the prediction results [10] ,[11].Using the feature selection algorithms accuracy can be enhanced, speed of the training process can be accelerated, storage cost can be decreased ,and also minimizes the dimensionality of the dataset [12] .In literature there have been various methods proposed to select the smaller number of features from the original feature set to get enhanced classification results. Those methods are feature extraction and feature selection [10] . When a new limited feature space can be derived from original feature space, it is called as feature extraction. On the other hand, in feature selection, some of the features can be selected from the original space without any transformation function.

Feature selection algorithms can be divided into two groups. Those are wrapper and filter. Wrapper function is on the basis of performance of specific classifier training. But this method maximizes the computational cost [13]. Whereas, the filter is on the basis of statistical characteristics of feature subset. There have been done a number of studies on feature selection to choose the best feature subset to accelerate the accuracy of classification technique. Information gain , ÷2 test, Euclidian distance, T-test, correlation based feature selection method, Markov blanket filter, fast correlation based feature selection are some of the common feature selection algorithms. However, this current study targets on applying a new approach based on information gain. It measures the weight or rank of an attribute by measuring the information gain with respect to the class.

$$\text{Information Gain (Class, Attribute)} = H(\text{Class}) - H(\text{Class} \mid \text{Attribute}).$$

In the literature, there are ensemble feature selection approaches also available. In [14], it was discussed and compared two different methods of combining feature selection and ensemble learning. First, Feature selection for ensemble learning. Second, Ensemble learning for feature selection. This method recorded predictive performance superior to traditional feature selection methods for supervised learning.

In article [13], the researchers proposed more robust feature selection techniques using an approach of ensemble feature selection techniques for high dimensional data. Authors of [15], applied a rotation forest ensemble decision tree algorithm and it is wrapped with best search technique. The wrapper uses forward selection to select the best feature subset dataset. The distinguished power of selecting features is evaluated using multiple machine learning techniques and the variety of the training data using the bagging algorithm. Two feature selection algorithms on the diabetic datasets are applied in [16] .First, feature selection via Supervise Model Construction then ReliefF . For analyzing the performance c4.5, Ib1 and naive Bayes classification algorithms were used and the performance of those algorithms are estimated. First algorithm recorded the improved performance on the classification than the second algorithm i.e. ReliefF.

Selective Naive Bayes is applied for filtering and ranking the feature of a medical dataset which gives the better performance by the researchers of [17]. The prior of single feature is calculated by adding the new feature to an existing set and comparing the performance of the current sets and previous sets. Dimensionality reduction of training and testing datasets have been proposed using novel Gaussian based kernel function by the authors of [18].

3. PROPOSED METHOD

The objective of our proposed method is to select the subset of features from huge feature space. If there is a requirement of selecting 'S' features of 'N' feature space, there have been different techniques available. Those techniques are Chi Squared Attribute Evaluator, Filtered Attribute Evaluator, and Gain Ratio Attribute Evaluator, Information Gain Attribute Evaluator. These techniques give the rank or weight to each attribute. According to the ranks given, top 'S' features will be selected. But, there is a chance of improving the performance if features are getting selected in different combinations.

In the proposed method, attribute ranking is calculated using information gain attribute evaluator. For searching method Ranker algorithm is applied. Ranker algorithm ranks the each attributes by their individual evaluations. After this process all the features or attributes are available in descending order of their ranks. After this, group the subset of features into clusters according to below given Figure.1 or Algorithm.

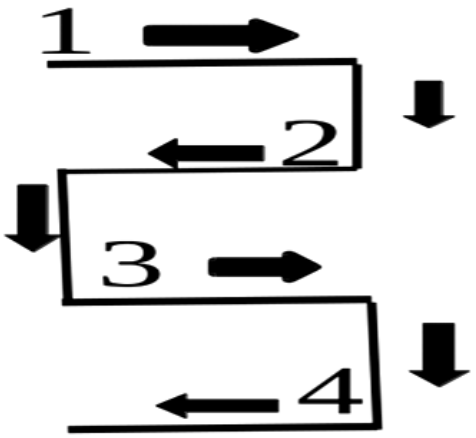

Figure 1: Cluster Format

Algorithm to form cluster of features

Input: N, S, K

Output: Subset of features

Let 'N' is Number of features
Let 'S' is Number of features to select
Let 'K' is the Number of clusters

Step 1: Find out the Rank of each Attribute

Step 2: Define number of features to select (S)

Step 3: Define number of clusters (K) = N/S

Step 4: Arrange First N/K features in descending order of Ranks from left to right in Level 1

Step 5: Arrange next N/K feature in descending order of Ranks from right to left in Level 2.

Step 6: Repeat Step 4 then step 5 for next Levels until the all features are arranged.

Step 7: Group, all vertically first order features of all levels in First cluster, Second order features of all levels in

A second cluster, and so on.

Note: In each cluster, N/K features will be organized, if N%S=0 (Zero), otherwise, few clusters(N%S) will have S+1 features.

Example:

Let N=15 (Number of features)

List of features according to their ranks are (A,B,C,D,E,F,G,H,I,J,K,L,M,N,O)

Let S=5 (Number of Clusters), So K=N/S=3

Form the cluster of features as per below table

Table -1 Cluster of features table

Level	Order No 1	Order No 2	Order No 3	Direction
1	A	B	C	Left to Right
2	F	E	D	Right to Left
3	G	H	I	Left to Right
4	L	K	J	Right to Left
5	M	N	O	Left to Right

Cluster 1 (K1) contains all Order No 1 features
K1={A,F,G,L,M}
Cluster 2 (K2) contains all Order No 2 features
K2={B,E,H,K,N}
Cluster 3 (K3) contains all Order No 3 features
K3={C,D,I,J,O}

4. EXPERIMENTS AND RESULTS

For the experiment of the proposed method, three datasets are used. The datasets are taken from http://archive.ics.uci.edu/ml. The description of datasets used in this study is given in Table 2.

Table 2
Cluster of features

Dataset	*# Features*	*# Instances*	*#Class*	*Type of Features*	*Remarks*
Chronic Kidney Disease	24	1170	2	Categorical, real	BalancedUsing SMOTE
Voting 1	16	435	2	Categorical, real	Imbalanced
Voting 2	16	2679	2	Categorical, real	BalancedUsing SMOTE

Initially chronic kidney disease dataset is not balanced, so these datasets are balanced by applying the SMOTE algorithm. After applying the SMOTE on initial dataset, total 1170 instances are generated. The same case is with Voting1 dataset. After applying the SMOTE on the initial voting1 dataset, total 2679 instances are generated .These dataset is named as Voting 2.

In the preprocessing stage InfoGainAttributeEval (Information Gain Attribute Evaluator) is applied to all the datasets separately to find out the rank of each feature. InfoGainAttributeEval used the Ranking algorithm to generate rank of each feature. The order of rankings in descending order is given in Table 3.

Table 3
Description of rank of features

Dataset	*# Features*	*Rank of features (Feature number is given as per rank)*
Chronic Kidney Disease	24	12,19,15,16,20,3,4,11,22,23,10,2,18,13,5,1,14,7,17,24,6,8,21,9
Voting 1	16	4,3,5,12,14,8,9,13,15,7,6,1,11,16,10,2
Voting 2	16	4,5,3,14,12,8,15,9,13,7,6,1,11,16,10,2

According to our proposed method described in section 3, we defined the features of clusters as mentioned in Table 4.

Table 4
Description of Number of clusters (k) and Number of features selected

Dataset Id	*Dataset*	*N*	*S*	*K =(N/S)*	*Cluster ID*	*Selected features in Cluster*
D1	Chronic Kidney Disease	24	8	3	KD11	12,3,4,2,18,7,17,9
					KD12	19,20,11,10,13,14,24,21
					KD13	15,16,22,23,5,1,6,8
D2	Chronic Kidney Disease	24	6	4	KD21	12,11,22,1,14,9
					KD22	19,4,23,5,7,21
					KD23	15,3,10,13,17,8
					KD24	16,20,2,18,24,6
D3	Voting 1	16	5*	3	KD31	4,8,9,1,11
					KD32	3,14,13,6,16
					KD33*	5,12,15,7,10,2
D4	Voting 2	16	4	4	KD41	4,9,13,2
					KD42	5,15,7,10
					KD43	3,8,6,16
					KD44	14,12,1,11

* One cluster will have 5+1=6 features, because N%S =1

N : # Features in original Dataset, S: # Features to select, K :# clusters (N/S)

10-fold cross validation is used to record the accuracy of various feature selector classifiers. In this experiment SMO (Sequential minimal optimization algorithm), Logistic regression model, Rule based algorithms like Jrip, OneR,Ridor are applied on the datasets w.r.t formed cluster of features described in Table 4 . The performance of each cluster of features is compared with Gain Ratio Attribute Evaluator (GRAE), Cfs Subset Evaluator(CfsE) .In the case of GRAE, top 'S' features are selected. In the case of CfsE, the features derived by the Attribute Subset Evaluator are selected. If the evaluator generates more than 'S' features, then CfsE is ensembled with GRAE and selected top 'S' features.

Classification accuracy is calculated using formula below

Accuracy = (TP+TN) / (TP+FN+FP+TN)

Where TP is # +ve records correctly classified.

TN is # -ve records correctly classified.

FN is # +ve records incorrectly classified as negative.

FP is # -ve records incorrectly classified as positive

Note: # means (Number of), +ve : positive , -ve negative

The performance of each cluster is described in tables below. Table 5. Describes the performance of clusters of Dataset D1. Table 6. Describes the performance of clusters of Dataset D2. Table 7. Describes the performance of clusters of Dataset D3. Table 8. Describes the performance of clusters of Dataset D4.

Table 5
Performance of clusters of Dataset D1

Cluster ID	*Jrip*	*OneR*	*Ridor*	*SMO*	*LR*
KD11 !	**98.71**	**93.93**	**97.17**	**96.92**	**97.69**
KD12 !	95.89	88.46	95.55	**92.82**	**94.44**
KD13 !	**97.77**	**94.61**	96.66	**97.52**	**98.71**
GRAE @	97.60	92.90	96.75	92.39	94.44
CfsE + GRAE @	97.60	92.90	97.26	92.30	94.44

Table 6
Performance of clusters of Dataset D2

Cluster ID	*Jrip*	*OneR*	*Ridor*	*SMO*	*LR*
KD21 !	96.23	**92.90**	95.98	87.09	92.22
KD22 !	**97.09**	92.05	96.23	**95.98**	**95.72**
KD23 !	**97.35**	**93.93**	**97.00**	**97.94**	**98.11**
KD24 !	**97.26**	**94.61**	96.75	**95.55**	**96.92**
GRAE @	97.09	92.90	96.97	92.47	94.70
CfsE @	97.26	92.90	97.26	92.30	94.87

Table 7
Performance of clusters of Dataset D3

Cluster ID	*Jrip*	*OneR*	*Ridor*	*SMO*	*LR*
KD31 !	**95.86**	**95.63**	**96.32**	**95.63**	**96.32**
KD32 !	87.35	87.35	86.20	87.81	88.27
KD33 !	86.20	83.67	86.89	87.12	86.20
GRAE @	95.63	95.63	95.63	95.63	95.63
CfsE @	95.40	95.63	94.71	95.63	96.09

Table 8
Performance of clusters of Dataset D4

Cluster ID	*Jrip*	*OneR*	*Ridor*	*SMO*	*LR*
KD41 !	**96.49**	**96.67**	96.60	**96.49**	**96.49**
KD42 !	88.05	86.74	87.94	87.98	87.98
KD43 !	87.79	87.45	87.86	87.57	87.45
KD44 !	87.30	85.18	88.24	88.50	88.91
GRAE @	96.49	96.67	96.75	96.64	96.49
CfsE @	97.12	96.67	96.71	96.49	96.41

Note:@ is existed method, ! Is our proposed method

According to the results generated, in every dataset, the subset of features of one or two clusters formed by our proposed method is improving the accuracy of one or two algorithms and few clusters are competing with algorithms, when compared with an existing Gain ratio attribute evaluator and Cfs. Those results are highlighted in the respective tables.

5. CONCLUSION AND FUTURE WORK

In this paper, we described a novel approach for selecting a subset of features using information gain. In this method, subset of features is formed in different clusters based on the algorithm described in section 3. All the features of clusters are analyzed on 3 different datasets in 4 different ways. All features of clusters are analyzed with Jrip, OneR, Ridor, SMO, LR and compared those results with existing techniques. From the dataset D1, cluster KD11 is improving the performance and KD13 is competing with existing techniques. From the dataset D2, cluster KD23 is improving the performance and KD22, KD24 are competing with existing techniques. From the dataset D3,cluster KD31 is performing better than existing techniques. From the dataset D4,cluster KD41 is competing with existing techniques. With the use of our proposed method, number of combinations to select a subset of features can be reduced ,once the rank of an attribute is derived from information gain.

In future, subset of feature selection methods using clusters can be applied in real time problems using ensembled feature selection and various searching techniques. Subset of feature selection is very much important in the case of Big Data, as large amount of data requires more computation and more storage. This can be reduced by a novel feature selection method and Map reduce programming model.

REFERENCES

[1] M. Sreedevi, G. V. Kumar, and L. S. S. Reddy, "Parallel and distributed approach for incremental closed regular pattern mining," in IT in Business, Industry and Government (CSIBIG), 2014 Conference on, 2014, pp. 1–5.

[2] H. Rahman, Ed., Data mining applications for empowering knowledge societies. Hershey, PA: Information Science Reference, 2009.

[3] D. Zhang, Y. Shi, Y. Tian, and M. Zhu, "A class of classification and regression methods by multiobjective programming," Frontiers of Computer Science in China, vol. 3, no. 2, pp. 192–204, 2009.

[4] E. L. Allwein, R. E. Schapire, and Y. Singer, "Reducing multiclass to binary: A unifying approach for margin classifiers," Journal of machine learning research, vol. 1, no. Dec, pp. 113–141, 2000.

[5] Z. Qi, Y. Tian, and Y. Shi, "A nonparallel support vector machine for a classification problem with universum learning," Journal of Computational and Applied Mathematics, vol. 263, pp. 288–298, Jun. 2014.

[6] J. D. Rennie and R. Rifkin, "Improving multiclass text classification with the support vector machine," 2001.

[7] S. Anbazhagan and N. Kumarappan, "A neural network approach to day-ahead deregulated electricity market prices classification," Electric Power Systems Research, vol. 86, pp. 140–150, May 2012.

[8] S. Prasad Potharaju and M. Sreedevi, "An Improved Prediction of Kidney Disease using SMOTE," Indian Journal of Science and Technology, vol. 9, no. 31, Aug. 2016.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of artificial intelligence research, vol. 16, pp. 321–357, 2002.

[10] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," Computers & Electrical Engineering, vol. 40, no. 1, pp. 16–28, Jan. 2014.

[11] V. Kumar, "Feature Selection: A literature Review," The Smart Computing Review, vol. 4, no. 3, Jun. 2014.

[12] M. Yao, M. Qi, J. Li, and J. Kong, "A novel classification method based on the ensemble learning and feature selection for aluminophosphate structural prediction," Microporous and Mesoporous Materials, vol. 186, pp. 201–206, Mar. 2014.

[13] Y. Saeys, T. Abeel, and Y. Van de Peer, "Robust feature selection using ensemble feature selection techniques," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2008, pp. 313–325.14.

[14] D. Guan, W. Yuan, Y.-K. Lee, K. Najeebullah, and M. K. Rasel, "A Review of Ensemble Learning Based Feature Selection," IETE Technical Review, vol. 31, no. 3, pp. 190–198, May 2014.

[15] A. Ozcift and A. Gulten, "A Robust Multi-Class Feature Selection Strategy Based on Rotation Forest Ensemble Algorithm for Diagnosis of Erythemato-Squamous Diseases," Journal of Medical Systems, vol. 36, no. 2, pp. 941–949, Apr. 2012.

[16] Y. Huang, P. McCullagh, N. Black, and R. Harper, "Feature selection and classification model construction on type 2 diabetic patients' data," Artificial Intelligence in Medicine, vol. 41, no. 3, pp. 251–262, Nov. 2007.

[17] T.-T. Wong and L.-H. Chang, "Individual attribute prior setting methods for naïve Bayesian classifiers," Pattern Recognition, vol. 44, no. 5, pp. 1041–1047, May 2011.

[18] G. R. Kumar, M. Nimmala, and G. Narasimha, "An Approach for Intrusion Detection Using Novel Gaussian Based Kernel Function," Journal of Universal Computer Science, vol. 22, no. 4, pp. 589–604, 2016.