2015 International Conference on Computing Communication Control and Automation

Experiments on Selection of pre-processing method for performance augmentation of Classifier

Katkar Vijay D., Almas A. K., Jaiswal Kajal
Department of Information Technology
Pimpri Chinchwad College of Engineering,
Pune, India

***Abstract*— Business and Research organizations are continuously generating huge amount of high dimensional data. They need to analyze this data in real-time with minimum cost. Data pre-processing techniques in combination with dimensionality reduction techniques are widely used by researchers to improve the quality of data and reduce the time, cost required to analyze the data. But standard methods are not available to select the combination of data pre-processing and feature selection techniques. This paper presents a novel method for selecting combination of pre-processing and feature selection method. Experimental results are provided to support the efficiency of proposed mechanism.**

Keywords—Data Mining, Data pre-processing, Feature selection

I. INTRODUCTION

Business and research organizations are generating huge amount of high dimensional data, which needs to be analyzed to improve the performance. But the cost, time of analysis is directly proportional to size and dimensionality of data. If the dimensionality of data is reduced, then it results in saving of time, cost. Combination of Data pre-processing techniques, feature selection techniques and Data mining algorithm is widely used by organizations and researchers for analysis of high dimensional data [1, 2, 3, 4].

Naïve Bayesian is one of the simplest algorithms and assumes that no two attributes are related with each other. From the given training set it calculates prior probability of each class. Then it calculates the likelihood probability of each class, multiplies both prior and likelihood probability to get posterior probability. This posterior probability is used to classify the records. It has high bias because of its prior assumption of independence. It has low variance as it is not too sensitive to changes in the data set. So it can be used with small as well as large data sets. To improve the efficiency of naïve bayes algorithm one should try to increase the number of attributes or change the structure of the data set. Increasing the number of instances won't help. It gives very good accuracy in cases where data sets increase exponentially with increase in attributes. It is not affected by unwanted attributes or outliners but affected by repeated data. It assumes numeric values have normal distribution.

J48 (C4.5) recursively selects an attribute to split the data set. Attribute that produces pure daughter node is considered for splitting. Information gain ratio of all the remaining attributes is calculated. An attribute with maximum gain ratio is selected as root (greedy approach).All the possible values of that attribute are taken as branches. This recursive process is continued. J48 has high variance because of its greedy approach and thus it is highly unstable. It can handle missing values by using proximity. It can easily handle numeric attributes. Local discretization of attributes takes place.

Random forest uses ensemble of classifiers (Randomization).In randomization from the original data set multiple data sets are produced by changing the number of attributes and keeping the instances same. Each data set is given input to different classifiers (stable as well as unstable).Output from each classifier is used for calculating the mode.

Random tree uses ensemble of classifiers (Bagging or Bootstrap). In bagging multiple data sets are produced by keeping attributes same and only changing the instances. These data sets are applied to unstable classifiers like decision tree. Average of different outputs of these classifiers for every bagged dataset is taken as a result. The accuracy of unstable classifiers increases due to decrease in variance. As the size of the data set given to each tree decreases variance also decreases.

Bayesian Networks (Bayes Net) are Directed Acyclic Graphs whose nodes represent variables and edges represent conditional dependency. Nodes that are not connected represent variables that are independent of other variables. Each node is associated with a probability function that takes as input a particular set of values for the node's

978-1-4799-6892-3/15 $31.00 © 2015 IEEE
DOI 10.1109/ICCUBEA.2015.94

parent's variables. It gives the probability of the variable represented by the node. Each variable is independent of its non descendants. It has more bias and less variance. It stores probabilities of its parents only, thus requires less memory space for storage.

PART is the enhanced version of C4.5 and RIPPER. It combines divide and conquer of C4.5 and separate and conquer of RIPPER. It does not produce large trees like decision tree. It is partial decision tree algorithm. Unlike C4.5, PART does not need global optimization (pruning). It works efficiently if the data is numerical and scaled.

Discretization PKI Discretization, Normalization, Numeric to Binary are widely used pre-processing techniques by many researchers [5, 6, 7, 8]. Discretization sorts the instances by its attributes value. After sorting, assign the value into ranges at the points where the class value changes. Numeric as well as string values can be discretized. It can use attribute values or class information.

PKI Discretization creates bins of equal frequency. Number of bins is equal to the square root of the number of missing values. It then replaces each bin with mean or median. It does not make use of class information. Therefore it is unsupervised.

Normalization tries to give all attributes an equal weight. It prevents attributes with initially large ranges from outweighing an attribute with small ranges. eg salary may outweigh some binary attribute. It is helpful in neural networks and distance measurement such as KNN. There are three types of normalizations: a) Min-Max normalization, b) Z-normalization, c) decimal scaling. But experiments show that Min_Max has less error.

Numeric to Binary uses Maximum and Minimum value of each numeric attributes to find the mid value. If the value of the numeric attribute is less than or equal to Mid value, the value of the new attribute is 0 else the value of the new attribute is 1.

II. LITERATURE SURVEY

K. R. Seeja [9] proposed 'closed frequent Itemset Mining' technique as a feature selection method for cancer classification and said that this technique works best with ELM and SVM classifiers. It used the maximum 26 attributes.

M Demetgul et al. [10] in his work on fault diagnosis of material handling system said that when LLE(Local Linear Embedding) was used with GK (Gustafson-Kessel) algorithm, accuracy was more than 90%.

Yue Huang et al. [11] in his work on type 2 diabetic patient's data used FSSMC feature selection with Naive Bayes, IB1 and C4.5 classifiers. Results of experiments showed that processing speed and performance of classifiers were improved.

P. Ravisankar et al. [12] in detection of financial statement fraud observed that PNN and GP outperformed other combinations of feature selection and classifier. But failed to observe that PNN and GP performed better with 18 features than with 10 features.

K. Rajeshwari et al. [13] in the paper for Ischemic Heart Disease Identification said that when Artificial Neural Network used as feature selector along with Feed Forward Neural Networks gives best accuracy on IHD dataset with 12 attributes. Accuracy reduces when attributes are further reduced.

Liuzhi Yin et al. [14] in paper ' Feature Selection for High Dimensional Imbalanced Data' said that performance becomes better with the increasing number of features is over certain value. There are accident drops when number of features increases. But the paper didn't specify any specific method to know that threshold value.

Xiangzhou Zhang et al.'s study [15] for stock prediction modeling conducted comparative experiments between CFL, PCA, DT, LASSO feature selection methods with different baseline model . CFS gave higher accuracy than others and proposed that CFS improve the performance of the stock prediction. But it was not observed that combination of CFS and any baseline classifier used maximum attributes.

Correlation based Feature Selection (CFS) is used feature selection method by researchers to reduce the dimensionality of data. It measures correlation between nominal features. It first discretizes the numeric features and then removes irrelevant and redundant attributes. CFS finds the feature-feature correlation and feature-class correlation using the formula:

$$R_{zc} = \frac{KR'_{zi}}{\sqrt{K + K(K-1)R'_{ii}}} \quad (1)$$

R_{zc}=corelation between summed components and outside variable

K=no of components

R'_{zi}=average of corelation between the components and outside variable

$\bar{R}_{ff}$=average inter-correlation between components

Bais of the evaluation function is towards subsets that contain features that are highly correlated with the class and less correlated with other features.

CFS uses three strategies:

a. RELIEF
b. MDL
c. Symmetric uncertainty

In most cases CFS performs well but it chooses irrelevant attribute when there are few training examples. A CFS filter works efficiently with large data sets. To prepare a final attribute set CFS applies a search method. This paper uses greedy stepwise searching method. Greedy stepwise can either use forward selection or backward elimination. In forward selection attributes are added one by one to empty set till it does it does not reduces the efficiency. In backward elimination attributes are deleted one by one.

III. PROPOSED MECHANISM

Normally first two steps of data analysis process are pre-processing and feature selection. Proposed mechanism is explained with the help of the figure 1.

It consists of 4 units:

a. Pre-processing unit
b. Feature selection unit
c. Classification unit
d. Analysis unit

Pre-processing unit: Data set is given as a input to pre-processing unit. It is responsible for applying various (Normalize, Discretize, PKI Discretize, Numeric to binary) pre-processing methods on it. Output of pre-processing unit is given as an input to Feature Selection unit.

Feature Selection unit: It receives pre-processed data from pre-processing unit and apply CFS method on it to select most relevant features. Then it eliminates less relevant features from data set. Dimensionally reduced data by this unit is given as a input to classification unit

Classification unit: This unit is responsible for applying specific data mining algorithm on received data. Results generated by applying data mining algorithm are given as input to Analysis unit.

Analysis unit: This unit determines which combination of pre-processing method, feature selection method is suitable for the problem domain.

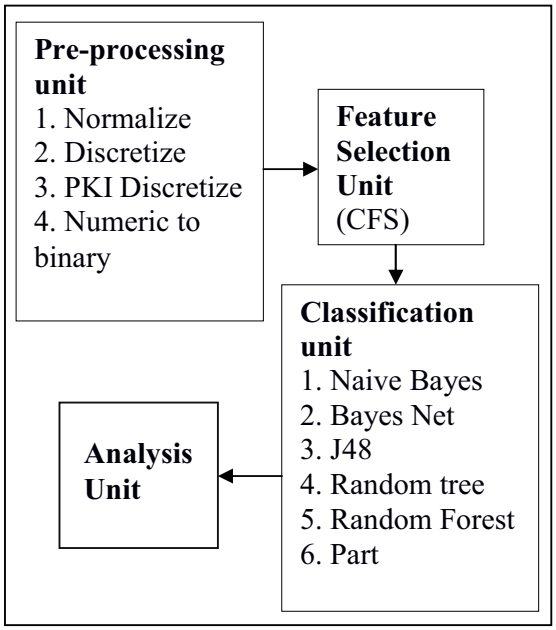

Fig. 1: Proposed Mechanism

IV. EXPERIMENTAL RESULTS

Experiments are performed on I3 machine (4GB RAM, 2.30 GHz processor) using Weka. Data sets used for the experiment are:

i.coil-2000

ii.credit-g

iii.bank

Dataset: coil-2000

Table 1, 2, 3, 4, 5, 6, 7 shows the experimental results obtained after applying Naïve Bayesian, Bayes Net, J48, Random Tree, Random Forest, REP Tree, PART algorithm respectively on coil-2000 dataset. Same file was used as training as well as testing dataset. It can be observed that maximum accuracy is achieved when number of features selected by feature selection algorithm is more. If less number of features are selected then detection accuracy is also less.

TABLE I: DETECTION ACCURACY OF NAÏVE BAYESIAN CLASSIFIER FOR COIL-2000 DATASET

Classifier	Pre-processing Method	No. of Attributes	Accuracy
Naive Bayes	Normalize	10	92.2012
	Discretize	7	93.0971
	PKI Disc.	9	92.4964
	N2B	10	93.6978

TABLE II: DETECTION ACCURACY OF BAYES NET CLASSIFIER FOR COIL-2000 DATASET

Classifier	Pre-processing Method	No. of Attributes	Accuracy
BayesNet	Normalize	10	92.7204
	Discretize	7	93.0971
	PKI Disc.	9	92.5168
	N2B	10	93.6876

TABLE III: DETECTION ACCURACY OF J48 CLASSIFIER FOR COIL-2000 DATASET

Classifier	Pre-processing Method	No. of Attributes	Accuracy
J48	Normalize	10	94.0338
	Discretize	7	94.0338
	PKI Disc.	9	94.0338
	N2B	10	94.0338

TABLE IV: DETECTION ACCURACY OF RANDOM TREE CLASSIFIER FOR COIL-2000 DATASET

Classifier	Pre-processing Method	No. of Attributes	Accuracy
Random Tree	Normalize	10	95.9581
	Discretize	7	94.1356
	PKI Disc.	9	94.9807
	N2B	10	94.2374

TABLE V: DETECTION ACCURACY OF RANDOM FOREST CLASSIFIER FOR COIL-2000 DATASET

Classifier	Pre-processing Method	No. of Attributes	Accuracy
Random Forest	Normalize	10	95.9071
	Discretize	7	94.1356
	PKI Disc.	9	94.9196
	N2B	10	94.2272

Dataset: credit-g

Table 7, 8, 9, 10, 11, 12, 13, 14 shows the experimental results obtained after applying Naïve Bayesian, Bayes Net, J48, Random Tree, Random Forest, REP Tree, PART algorithm respectively on credit-g dataset. Same file was used as training as well as testing dataset. It can be observed that maximum accuracy is achieved when number of features selected by feature selection algorithm is more. If less number of features are selected then detection accuracy is also less.

TABLE VI: DETECTION ACCURACY OF REP TREE CLASSIFIER FOR COIL-2000 DATASET

Classifier	Pre-processing Method	No. of Attributes	Accuracy
REP Tree	Normalize	10	94.0338
	Discretize	7	94.0338
	PKI Disc.	9	94.0338
	N2B	10	94.0338

TABLE VII: DETECTION ACCURACY OF PART CLASSIFIER FOR COIL-2000 DATASET

Classifier	Pre-processing Method	No. of Attributes	Accuracy
PART	Normalize	10	94.0338
	Discretize	7	94.0338
	PKI Disc.	9	94.0338
	N2B	10	94.0338

TABLE VIII: DETECTION ACCURACY OF NAÏVE BAYESIAN CLASSIFIER FOR CREDIT-G DATASET

Classifier	Pre-processing Method	No. of Attributes	Accuracy
Naive Bayes	Normalize	3	74.2
	Discretize	5	75.5
	PKI Disc.	4	75.1
	N2B	4	72.3

TABLE IX: DETECTION ACCURACY OF BAYES NET CLASSIFIER FOR CREDIT-G DATASET

Classifier	Pre-processing Method	No. of Attributes	Accuracy
BayesNet	Normalize	3	74.9
	Discretize	5	75.5
	PKI Disc.	4	75.1
	N2B	4	72.3

TABLE X: DETECTION ACCURACY OF J48 CLASSIFIER FOR CREDIT-G DATASET

Classifier	Pre-processing Method	No. of Attributes	Accuracy
J48	Normalize	3	75.2
	Discretize	5	76.7
	PKI Disc.	4	74.8
	N2B	4	74.3

TABLE XI: DETECTION ACCURACY OF RANDOM TREE CLASSIFIER FOR CREDIT-G DATASET

Classifier	Pre-processing Method	No. of Attributes	Accuracy
Random Tree	Normalize	3	80
	Discretize	5	81.3
	PKI Disc.	4	84.7
	N2B	4	74.7

TABLE XII: DETECTION ACCURACY OF RANDOM FOREST CLASSIFIER FOR CREDIT-G DATASET

Classifier	Pre-processing Method	No. of Attributes	Accuracy
Random Forest	Normalize	3	79.9
	Discretize	5	81.3
	PKI Disc.	4	84.4
	N2B	4	74.6

TABLE XIII: DETECTION ACCURACY OF REP TREE CLASSIFIER FOR CREDIT-G DATASET

Classifier	Pre-processing Method	No. of Attributes	Accuracy
REP Tree	Normalize	3	73.7
	Discretize	5	75.4
	PKI Disc.	4	77
	N2B	4	74.3

TABLE XIV: DETECTION ACCURACY OF PART CLASSIFIER FOR CREDIT-G DATASET

Classifier	Pre-processing Method	No. of Attributes	Accuracy
PART	Normalize	3	74.2
	Discretize	5	75.5
	PKI Disc.	4	75.1
	N2B	4	72.3

Dataset: bank

Table 15, 16, 17, 18, 19, 20, 21 shows the experimental results obtained after applying Naïve Bayesian, Bayes Net, J48, Random Tree, Random Forest, REP Tree, PART algorithm respectively on bank dataset. Same file was used as training as well as testing dataset. It can be observed that maximum accuracy is achieved when number of features selected by feature selection algorithm is more. If less number of features are selected then detection accuracy is also less.

TABLE XV: DETECTION ACCURACY OF NAÏVE BAYESIAN CLASSIFIER FOR BANK DATASET

Classifier	Pre-processing Method	No. of Attributes	Accuracy
Naive Bayes	Normalize	4	89.3386
	Discretize	4	89.4714
	PKI Disc.	5	89.6262
	N2B	4	89.2502

TABLE XVI: DETECTION ACCURACY OF BAYES NET CLASSIFIER FOR BANK DATASET

Classifier	Pre-processing Method	No. of Attributes	Accuracy
BayesNet	Normalize	4	89.4714
	Discretize	4	89.4714
	PKI Disc.	5	89.7589
	N2B	4	89.2502

TABLE XVII: DETECTION ACCURACY OF J48 CLASSIFIER FOR BANK DATASET

Classifier	Pre-processing Method	No. of Attributes	Accuracy
J48	Normalize	4	89.4714
	Discretize	4	89.4714
	PKI Disc.	5	89.7589
	N2B	4	89.2502

TABLE XVIII: DETECTION ACCURACY OF RANDOM TREE CLASSIFIER FOR BANK DATASET

Classifier	Pre-processing Method	No. of Attributes	Accuracy
Random Tree	Normalize	4	89.4714
	Discretize	4	89.4714
	PKI Disc.	5	89.7589
	N2B	4	89.2502

TABLE XIX: DETECTION ACCURACY OF RANDOM FOREST CLASSIFIER FOR BANK DATASET

Classifier	Pre-processing Method	No. of Attributes	Accuracy
Random Forest	Normalize	4	89.4714
	Discretize	4	89.4714
	PKI Disc.	5	89.7589
	N2B	4	89.2502

TABLE XX: DETECTION ACCURACY OF REP TREE CLASSIFIER FOR BANK DATASET

Classifier	Pre-processing Method	No. of Attributes	Accuracy
REP Tree	Normalize	4	89.4714
	Discretize	4	89.4714
	PKI Disc.	5	89.7589
	N2B	4	89.2502

TABLE XXI: DETECTION ACCURACY OF PART CLASSIFIER FOR BANK DATASET

Classifier	Pre-processing Method	No. of Attributes	Accuracy
PART	Normalize	4	89.4714
	Discretize	4	89.4714
	PKI Disc.	5	89.7589
	N2B	4	89.2502

V. CONCLUSION

This paper presents a novel method for selection of proper combination of pre-processing and feature selection method. It can be observed from experimental results that, the combination of pre-processing and feature selection method that results in maximum number of selected features should be used to improve the data classification efficiency. If combination which gives less number of features is used then it may affect the accuracy of the classifier.

REFERENCES

[1] Liang Zhang, Lingling Zhang, Weili Teng, Yibing Chen, "Based on Information Fusion Technique with Data Mining in the Application of Finance Early-Warning", Procedia Computer Science 17 (2013) 695 – 703

[2] Shu-Hsien Liao, Shan-Yuan Chou, "Data mining investigation of co-movements on the Taiwan and China stock markets for future investment portfolio", Expert Systems with Applications 40 (2013) 1542–1554

[3] Chia-Ming Wang, Yin-Fu Huang, "Evolutionary-based feature selection approaches with new criteria for data mining: A case study of credit approval data", Expert Systems with Applications 36 (2009) 5900–5908

[4] Sebastián Maldonado, Richard Weber, Fazel Famili, "Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines", Information Sciences 286 (2014) 228–246

[5] Weijun li, Zhenyu Liu, "A method of SVM with Normalization in Intrusion Detection", Procedia Environmental Sciences 11 (2011) 256 – 262

[6] Jonathan J. Davis, Andrew J. Clark, "Data preprocessing for anomaly based network intrusion detection: A review", computers & security 30 (2011) 353- 375

[7] D.H. Pandya, S.H. Upadhyay, S.P. Harsha, "Fault diagnosis of rolling element bearing with intrinsic mode function of acoustic emission data using APF-KNN", Expert Systems with Applications 40 (2013) 4137–4145

[8] Diego Garc´ıa-Saiz, Marta Zorrilla, "Towards the development of a classification service for predicting students' performance", The 6th International Conference on Educational Data Mining, EDM 2013, 318-319

[9] K.R. Seeja, "Feature selection based on closed frequent itemset mining: A case study on SAGE data classification", Neurocomputing (2014),http://dx.doi.org/10.1016/j.neucom.2014.03.084

[10] M. Demetgul, K. Yildiz, S. Taskin, I. N. Tansel, O. Yazicioglu, "Fault diagnosis on material handling system using feature selection and data mining techniques", Measurement 55 (2014) 15–24

[11] Yue Huang, Paul McCullagh, Norman Black, Roy Harper, "Feature selection and classification model construction on type 2 diabetic patients' data", Artificial Intelligence in Medicine (2007)41, 251—262

[12] P. Ravisankar, V. Ravi, G. Raghava Rao, I. Bose, "Detection of financial statement fraud and feature selection using data mining techniques", Decision Support Systems 50 (2011) 491–500

[13] K.Rajeswari, Dr. V. Vaithiyanathan, Dr. T.R. Neelakantan, "Feature Selection in Ischemic Heart Disease Identification using Feed Forward Neural Networks", Procedia Engineering 41 (2012) 1818 – 1823

[14] Liuzhi Yin, Yong Ge, Keli Xiao, Xuehua Wang, Xiaojun Quan, "Feature selection for high-dimensional imbalanced data", Neurocomputing 105 (2013) 3–11

[15] Xiangzhou Zhang, Yong Hu, Kang Xie, Shouyang Wang, E.W.T. Ngai, Mei Liu, "A causal feature selection algorithm for stock prediction modelling", Neurocomputing 142 (2014) 48–59