International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:6, No:2, 2012

Decision Tree-based Feature Ranking using Manhattan Hierarchical Cluster Criterion

Yasmin Mohd Yacob, Harsa A. Mat Sakim and Nor Ashidi Mat Isa

***Abstract*—**Feature selection study is gaining importance due to its contribution to save classification cost in terms of time and computation load. In search of essential features, one of the methods to search the features is via the decision tree. Decision tree act as an intermediate feature space inducer in order to choose essential features. In decision tree-based feature selection, some studies used decision tree as a feature ranker with a direct threshold measure, while others remain the decision tree but utilized pruning condition that act as a threshold mechanism to choose features. This paper proposed threshold measure using Manhattan Hierarchical Cluster distance to be utilized in feature ranking in order to choose relevant features as part of the feature selection process. The result is promising, and this method can be improved in the future by including test cases of a higher number of attributes.

***Keywords*—**Feature ranking, decision tree, hierarchical cluster, Manhattan distance.

I. Introduction

THE goal of feature selection is to reduce dimensionality and find essential features to improve classifier accuracy. In addition, feature selection can reduce classifier processing load and time, thus resulted in cost-effective processing. In order to find important features, the previous feature selection algorithm utilized various searching methods. This searching method includes forward, backward and floating technique [1]. Heuristic search method which includes feature ranking is also one of the techniques to search essential features. In the feature ranking scope, one can find essential features via information theoretic criterion namely mutual information (MI) [2][3][4][5], correlation-based measure namely, symmetrical uncertainty (SU) [6][7][8][9] and decision tree [10][11][12]. Mutual information determines how much one random variable tells about another variable. The computed mutual information value generated the ranking of importance. Battiti selected relevant features using MI ranking with specified threshold value [2]. On the other hand, Fleuret proposed feature selection method based on conditional mutual information [3].

Yasmin Mohd Yacob is with the School of Electrical and Electronic Engineering, USM, Malaysia, phone: +6(04) 5995096; (e-mail: ymy.ld09@student.usm.my).

Harsa Amylia Mat Sakim is with the School of Electrical and Electronic Engineering, USM, Malaysia, phone: +6(04) 5995821; (e-mail: amylia@eng.usm.my).

Nor Ashidi Mat Isa is with the School of Electrical and Electronic Engineering, USM, Malaysia, phone: +6(04) 5996051; (e-mail: ashidi@eng.usm.my).

Wren proposed Mutual Information Measure (MIM) to choose features and implemented shared minimum MIM to compute strength of association between features, which actually acted as the threshold measure [4]. Peng et al. developed mutual information-based minimal-redundancy-maximal-relevance (mRMR) method as a measure to select essential features in feature ranking [5]. Indeed, most mutual information-based feature selection methods were based on the feature ranking notion. Next, feature ranking studies implemented symmetrical uncertainty that measured the correlation between features and the class itself [13]. Both studies by Chou et al. and Kannan et al. proposed feature ranking via symmetrical uncertainty between feature and class, and feature and feature [6][9]. Whereas, Piroonratana et al. developed Round Robin symmetrical uncertainty ranking by considering SU between feature and class [8]. Indeed, Osman et al. developed Correlation Forward Selection (CorrFS) by means of correlation ranking measure [7]. In fact, there were actually many feature rankings researches utilizing SU criterion [14][15][16]. On the other hand, there were not so many studies on decision tree-based feature selection. Tao-Wang et al. supported in his paper that the impact of feature selection in decision tree learning was not well studied [17]. In feature selection studies, it is important to identify the relevant features first. This is because from the relevant features, redundant features can be identified and removed with further processing via additional metric. Some of the decision tree-based feature selection studies were by Zhou et al. who proposed a statistical–heuristic feature selection criterion [10]. His method does not implement ranking mechanism, instead generated the tree and remove irrelevant features by utilizing pruning condition, namely symmetrical tau. In his study, the pruning condition acted as the threshold mechanism to select relevant features. The pruning condition was developed from the Goodman and Kruskal measure of association. The measure of association actually resembled the concept of symmetrical uncertainty.

On the other hand, Hwang et al. study was quite similar with the proposed method. He developed a decision tree-based features ranking via Information Gain Ratio (IGR) metric, which resulted a partially ordered set of features. Selection of relevant features was determined from descending ordered features with a threshold measure, $\mu + k\sigma$. Hwang's threshold measure revolved around the mean and standard deviation of the IGR value, and the user needed to consider the correct coefficient k to determine relevant features [11].

The next decision tree-based feature selection study was developed by Mohammadi et al., which utilized Gini split criteria on the decision tree. Balanced error rate (BER) for both the tree and pruned tree was computed. Then, node importance index was calculated while the tree is not empty. Node's importance index was determined from the difference between BER of the pruned tree and the original tree. The process will return the list of important features with threshold value of the node important index approaching value of 1 [12]. It was observed that even within the decision tree-based feature selection studies, some studies implemented a straight forward ranking method with measures of threshold, and some generated the tree and the pruning criteria acted as the threshold measures to select relevant features. Regardless of the differences, this paper focused on decision tree-based feature selection method in order to choose essential features. In fact, Ratanamahatana and Gunopulos also demonstrated the usage of the decision tree as the feature selector [18]. This paper proposes a threshold measure in the decision tree-based feature selection to choose relevant features. The threshold measure is based on the combination of mutual information metric and adopting Manhattan distance measures from the hierarchical clustering notion. It promises a simple yet effective method in order to select relevant features.

II. Hierarchical Cluster

Hierarchical clustering is a method of cluster analysis, which is applied to develop a hierarchy of clusters. There are two types of approaches to build hierarchical clusters, which are by means of agglomerative or bottom-up approach and divisive or top-down approach. A measure of dissimilarity between sets of observations is required to determine which clusters should be merged in the bottom-up scheme or where a cluster should be split in the top-down scheme. The measure of dissimilarity is derived from the measure of distance between pairs of observations, and a linkage criterion or condition that determines the dissimilarity as a function of the pair wise distances of observations in the sets. Some of the metric available in clustering are Euclidean distance, Manhattan distance, Mahalanobis distance and several others. Some commonly used linkage criteria is maximum or complete linkage clustering, minimum or single linkage cluster, mean or average linkage clustering and others [19].

III. The Proposed Manhattan Hierarchical Cluster Threshold Criterion

It is mentioned from the earlier section that the searching of important features is carried out by heuristic searching via decision tree induction. The proposed threshold criterion follows the same search method but instead uses information theoretic metric namely Gini Gain. When a particular feature with maximum value of Gini Gain was chosen, the feature will not be included in the next iteration and the group of instances from other features that is adjacent to the leaf node from the maximum feature are also reduced. At the end of the iteration, it resulted in ranks of features but without a threshold that separates features from the irrelevant ones. This paper proposes a threshold metric which results in the selection of relevant features based on Manhattan hierarchical cluster. Despite the feature ranking generated from the decision tree induction, mutual information is utilized in the hierarchical cluster distance computation. This is because, according to Principe et al. and Torkkola et al., mutual information is proven to be optimal to perform class separation [20][21]. Thus, the feature ranking is a decision tree-based but integrated with mutual information in determining distance between the hierarchy of clusters. On the other hand, the relationship between the ranked features can be viewed as features in hierarchical clusters. Figure 1 illustrates the proposed notion. Previously, it is mentioned that the cluster dissimilarity metric determines which cluster should be combined or split. Thus, features in a hierarchical cluster can integrate the cluster dissimilarity concept to determine the threshold to choose relevant features. The Manhattan distance is proposed in the study due to suitability based on heuristic analysis of the background study. The proposed Manhattan distance measure is described in (1) and Table 1 demonstrates the Manhattan distance equation computation. The proposed Manhattan hierarchical cluster criterion accumulated mutual information difference between features in a hierarchy. It is also mentioned earlier that besides the distance measure, the linkage criterion or condition is essential to determine the effectiveness of dissimilarity measure in the hierarchical clusters. The threshold condition is determined from the slope of the cubic polynomial equation based on the Manhattan hierarchical cluster value. If the slope is 0 or negative, it showed the features from the point onwards are irrelevant. If the slope is positive and keep inclining, all the features are possibly relevant.

$$\text{Proposed Manhattan distance} = \sum_{i=1}^{n} \left(a_i - b_i \right) \quad (1)$$

$$y = ax^3+bx^2+cx+d \quad (2)$$

Equation (2) describes the cubic polynomial equation whereby the coefficient a, b, c and d are determined from running the curve fitting tool in Matlab. The proposed threshold condition is derived from heuristic analysis of eight data sets namely Iris, Thyroid, Pima Indian Diabetes, Breast Cancer, Wine, including Monk data set, which is the benchmark data set to determine relevant features. Figure 2-7 illustrate how the threshold condition is derived.

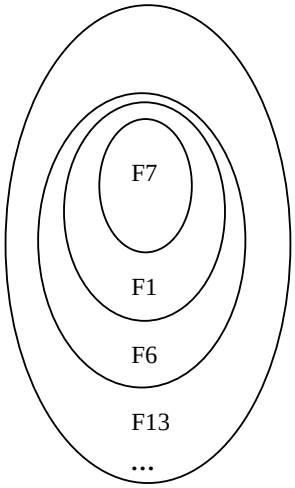

Fig. 1 Hierarchical clusters of features

TABLE I
PROPOSED MANHATTAN HIERARCHICAL CLUSTER DISTANCE MEASURE

Ranked features	MI value	Difference	Manhattan
F7	a	a-b	a-b
F10	b	b-c	(a-b) + (b-c)
F6	c	c-d	(a-b)+(b-c)+(c-d)
F13	d	…	…
…	…	…	…

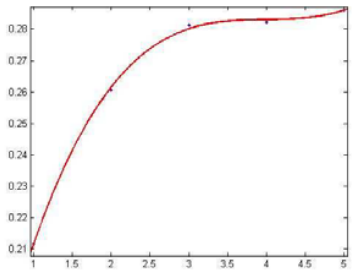

Fig. 2 Derivation of Monk1 data set threshold condition

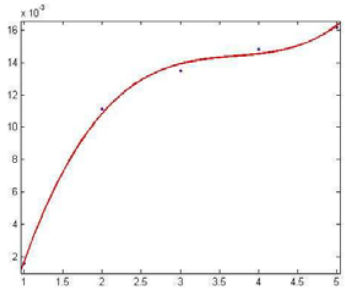

Fig. 3 Derivation of Monk2 data set threshold condition

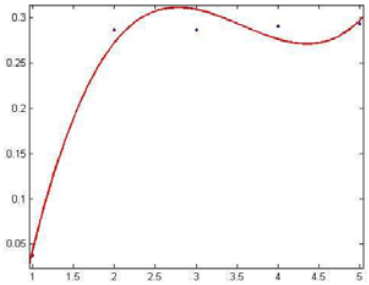

Fig. 4 Derivation of Monk3 data set threshold condition

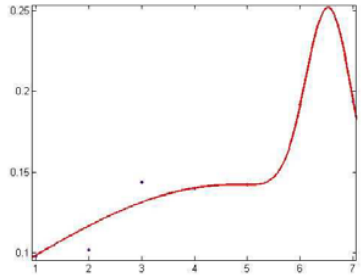

Fig. 5 Derivation of Pima Indian Diabetes data set threshold condition

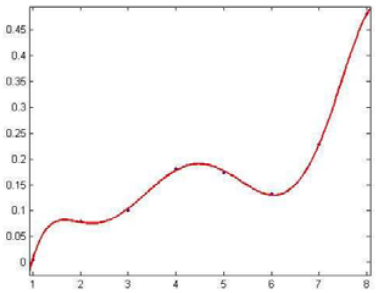

Fig. 6 Derivation of Breast Cancer data set threshold condition

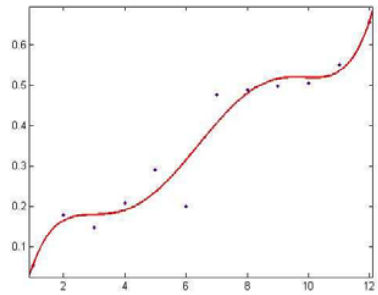

Fig. 7 Derivation of Wine data set threshold condition

IV. EXPERIMENTS

The proposed criterion are tested using eight data sets taken from the UCI machine learning repository, including the Monk data set, which act as the benchmark [22]. The data sets information is demonstrated in Table II. The number of

instances for the study ranges from 122 to 768 with the number of attributes up to 13 attributes. The analysis is done among decision tree-based feature selection methods. Iris, Thyroid, Pima Indian Diabetes, Breast Cancer and Wine are continuous data sets that have been discretized using Hybrid Dynamic Window Pairwise Gini (HDWPG) discretization algorithm [23]. The data sets need to be discretized because the decision tree utilized information theoretic metric to determine the important features. The results are classified using Multilayer Perceptron (MLP) from WEKA software package [24].

V. Results And Discussion

The results are evaluated based on the benchmark data set, accuracy and statistical tests. According to Liu and Motoda, relevant features from the Monk1 data set are 3 out of six features. Whereas for the Monk2 data set, all the features are relevant. Meanwhile, 3 out of six features from the Monk3 data set are relevant features [25]. Based on Table III and Table IV, the proposed decision tree-based feature ranking using Manhattan Hierarchical Cluster threshold criterion (MHCC) satisfies the benchmark requirement of selecting relevant features from the Monk data set. Even though Zhou and Mohammadi passed Monk3 and Monk2 benchmark test respectively, it can be concluded that Zhou, Hwang and Mohammadi methods are less likely conformed to the benchmark cases.

Results that show the relevant features and number of features for each data set are described in Table III and Table IV. Table V and Table VI assessed the decision tree-based feature selection methods effectiveness in terms of accuracy and statistical test. From Table IV, it can be concluded that Mohammadi's method generated the least number of features, however the result of error rate or root mean squared-error (RMSE) from Table V showed poor results whereby it only wins 1 out of eight cases.

Meanwhile, Zhou's and Hwang's method seemed to generate a similar number of features. However, in terms of error rate showed in Table V, Hwang's technique performed much better than Zhou's method with 3 out of eight winnings.

This showed that Hwang produced a better selection of features compared to Zhou. Manhattan Hierarchical Cluster criterion (MHCC) wins 3 out of eight cases which indeed comparable to Hwang's work.

This paper also attempts to assess the effectiveness of the proposed threshold measure using statistical test. According to Powers and Xie, evaluation on the selection of features needed to be done using Binary Logistic Regression for binary classes and Multinomial Logistic Regression for 3 or more classes [26]. Both statistical tests were chosen because the data sets are categorical data. The features selected from the regression tests were based on 0.05 significant p-value. By comparing results from Table VI with Table III and Table IV, it showed that Zhou's work is comparable with the statistical test in one out of eight cases. The data set that adheres to the statistical test is Iris. On the other hand, 4 out of eight cases from Hwang's study correspond to statistical test. The data sets involved are Iris, Thyroid, Monk1 and Monk3. In contrast, Mohammadi's method does not really conform to the statistical test. Meanwhile, MHCC is comparable to the statistical tests in one out of eight cases, which is the Breast Cancer data set. As a result, statistical test is a less suitable mechanism to verify the selection of relevant features from a decision tree.

The selection of features from the statistical test in Table VI specifically from the Monk data set does not conform to the benchmark requirement mentioned in Liu and Motoda [24]. This is because the statistical test resulted, Feature 1 and Feature 5 from the Monk1 data set as two relevant features whereas Liu and Motoda claimed there are three relevant features from the Monk 1 data set. Whereas, the statistical test selected one feature as important when Liu and Motoda claimed all the features in Monk2 data set are relevant. Liu and Motoda claimed Monk3 data set has three relevant features whereas the statistical test listed two features and failed to conform to the benchmark test. The discrepancy is maybe because the statistical test has determined its significant features and discards irrelevant features whereas Liu and Motoda considered the relevant features only. Regardless of the discrepancy in the assessment part, the proposed Manhattan Hierarchical Cluster criterion can be an effective measure to select relevant features due to comparable error rate. The threshold criterion implements a heuristically simple yet powerful method in order to select relevant features. Nevertheless, the method can be improved in the future by including test cases of a higher number of attributes.

TABLE II
DATA SETS INFORMATION

Dataset	Instances	Attribute	Class
Iris	150	4	3
Thyroid	215	5	3
Diabetes	768	8	2
Breast Cancer	699	9	2
Wine	178	13	3
Monk1	124	6	2
Monk2	169	6	2
Monk3	122	6	2

TABLE III
RELEVANT FEATURES FROM ZHOU AND HWANG DECISION TREE-BASED FEATURE SELECTION

Dataset	Without Feature Selection	With Feature Selection			
		Zhou		Hwang	
		Number of features	Relevant features	Number of features	Relevant features
Iris	4	2	4,3	2	3,4
Thyroid	5	4	4,2,3,5	3	2,3,5
Diabetes	8	7	7,6,4,8,5,2,3	4	2,5,6,8
Breast Cancer	9	6	4,5,6,7,8,9	4	2,3,6,8
Wine	13	8	5,6,7,9,10,11,12,13	6	1,7,10,11,12,13
Monk1	6	4	1,3,4,5	2	5,1
Monk2	6	3	3,4,5	3	4,5,6
Monk3	6	3	5,1,2	2	2,5

TABLE IV
RELEVANT FEATURES FROM MOHAMMADI AND MHCC DECISION TREE-BASED FEATURE SELECTION

Dataset	With Feature Selection			
	Mohammadi		MHCC	
	Number of features	Relevant features	Number of features	Relevant features
Iris	1	4	3	4,1,3
Thyroid	1	4	4	4,1,3,2
Diabetes	3	2,8,1	6	2,8,6,1,7,4
Breast Cancer	2	6,2	9	2,3,6,7,1,8,5,4
Wine	2	1,11	13	7,10,6,13,12,11,1,9,2,8,4,5,3
Monk1	5	2,5,1,6,4	3	5,1,4
Monk2	6	1,3,6,2,5,4	6	5,4,6,1,2,3
Monk3	4	5,2,3,4	3	2,5,1

TABLE V
ROOT MEAN SQUARED ERROR (RMSE) AMONG COMPARED METHODS

Data set	Zhou	Hwang	Mohammadi	MHCC
Iris	0.1559	0.1559	0.1562	**0.1167**
Thyroid	0.2701	**0.2562**	0.5514	0.2851
Diabetes	0.4834	**0.4303**	0.4553	0.4697
Breast Cancer	0.4842	0.4479	0.4512	**0.4315**
Wine	0.2561	**0.1765**	0.6751	0.1829
Monk1	0.5348	0.4862	**0.4465**	0.4956
Monk2	**0.5245**	0.5258	0.5659	0.5659
Monk3	0.3325	0.3336	0.2937	**0.3325**

TABLE VI
IMPORTANT FEATURES VIA STATISTICAL TEST

Data set	Number of features	List of features
Iris	2	F3,F4
Thyroid	3	F3,F4,F5
Diabetes	5	F1,F2,F3,F6,F7
Breast Cancer	7	F1,F2,F3,F4,F6,F7,F8
Wine	7	F1,F4,F6,F7,F10,F12,F13
Monk1	2	F1,F5
Monk2	1	F4
Monk3	2	F2,F5

VI. Conclusion

In this paper, An efficient threshold criterion to select relevant features in decision tree-based feature selection has been introduced by integrating Manhattan hierarchical cluster notion. Based on the result, the proposed method gives promising results. The computation time and load can be reduced and concurrently may improve the accuracy of the classification process due to selection of relevant features. In conclusion, the proposed method is promising but need improvement in the measure and consider data sets with a higher number of attributes.

Acknowledgment

Part of this work is supported by USM Research University Grant no. 814082 and Universiti Sains Malaysia Research University Postgraduate Research Scheme grant (No. 1001/PELECT/8042009)

References

[1] P. Pudil, Novovicova, and J. Kittler, "Floating search methods in feature selection," Pattern Recognition Letters, vol. 15, pp. 1119-1125, 1994.
[2] R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Network Learning," IEEE Transactions on Neural Network,, pp. 537-550, 1994.
[3] F. Fleuret, "Fast Binary Feature Selection with Conditional Mutual Information," Journal of Machine Learning Research , vol. 5, pp. 1531-1555, 2004.
[4] J. Wren, "Extending the mutual information measure to rank inferred literature relationships," BMC Bioinformatics, pp. 145, 2004.
[5] H. Peng, F. Long, and C. Ding, "Feature Selection Based on Mutual Information:Criteria of Max-Dependency, Max-Relevance,and Min-Redundancy," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1226-1238, 2005.
[6] T.-S. Chou, K.K. Yen, J. Luo, N. Pissinou, and K. Makki, "Correlation-Based Feature Selection for Intrusion Detection Design," in IEEE 2007 Proc. Military Communications Conference (MILCOM 2007) , pp. 1-7.
[7] H.E. Osman, "Correlation-based feature ranking for online classification," in IEEE 2009 International Conference on Systems, Man and Cybernetics (SMC 2009), pp. 3077-3082.
[8] T. Piroonratana, W. Wongseree, T. Usavanarong, A. Assawamakin, C. Limwongse, and N. Chaiyaratana, "Identification of Ancestry Informative Markers from Chromosome-Wide Single Nucleotide Polymorphisms Using Symmetrical Uncertainty Ranking," in 20th International Conference on Pattern Recognition (ICPR), 2010, pp. 2448-2451.
[9] S. Senthamarai Kannan, and N. Ramaraj, "A novel hybrid feature selection via Symmetrical Uncertainty ranking based local memetic search algorithm," Knowledge-Based Systems, pp. 580-585, 2010.
[10] X.J. Zhou, T.S. Dillon, "A statistical-heuristic feature selection criterion for decision tree induction," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 13, no. 8, pp. 834-841.
[11] Y-S Hwang, Hae-Chang R. "Decision tree decomposition-based complex feature selection for text chunking," in Proceedings of the 9th International Conference on Neural Information Processing (ICONIP '02), 2002, vol.2215, pp. 2217-2222.
[12] Mohammadi M, Gharehpetian GB. "Application of core vector machines for on-line voltage security assessment using a decisiontree-based feature selection algorithm," Generation, Transmission & Distribution, IET, vol. 3, no. 8, pp. 701-712, 2009
[13] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling, Numerical recipes in C, Cambridge University Press, 1988.
[14] L. Yu, and H. Liu, "Feature Selection for High-Dimensional Data : A Fast Correlation-Based Filter Solution," in Proceedings of the 20th International Conference on Machine Learning (ICML), 2003, pp. 856-863.
[15] B. Jiang, X. Ding, L. Ma, Y. He, T. Wang, and W. Xie, "A hybrid feature selection algorithm: Combination of symmetrical uncertainty and genetic algorithms", in Proc. of the 2nd. Intl. Symposium on Optimization and Systems Biology (OSB'08), Lijang, China, 2008, pp. 152-157.
[16] M.A. Hall, and L.A. Smith, "Feature Selection for Machine Learning : Comparing A Correlation-based Filter Approach to the wrapper," in Proceedings of the 12th International Florida Artificial Intelligence Research Society Conference, 1999, pp. 235-239.
[17] Tao-Wang, Zhen xing Qin, Zhi Jin Shichao Zhang, "Handling overfitting in test cost-sensitive decision tree learning by feature selection, smoothing and pruning", Journal of Systems and Software, vol. 83, pp. 1137-1147, 2010.
[18] C. Ratanamahatana, and D. Gunopulos, "Feature Selection for the Naive Bayesian Classifier using Decision Trees," Applied Artificial Intelligence, pp. 475-487, 2003.
[19] G. Gan, C. Ma, and J. Wu, "Data clustering : theory, algorithms, and applications", Philadelphia, Pa.: Society for Industrial and Applied Mathematics, 2007.
[20] J.C. Principe, J.W. Fisher III, and D. Xu. "Information theoretic learning". In Simon Haykin, editor, Unsupervised Adaptive Filtering. Wiley, New York, NY, 2000.
[21] K. Torkkola. "Feature extraction by non-parametric mutual information maximization," Journal of Machine Learning Research, vol. 3, pp. 1415–1438, 2003.
[22] Frank, A. & Asuncion, A. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2010.
[23] Yasmin Mohd Yacob, Harsa A. Mat Sakim and Nor Ashidi Mat Isa, "A Hybrid Discretization Algorithm Based on Dynamic Window and Hybrid Gini Criterion," Data Mining and Knowledge Discovery, submitted for publication.
[24] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, The WEKA Data Mining Software: An Update, SIGKDD Explorations, vol. 11, No. 1, 2009.
[25] H. Liu, and H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, Springer, 1998.
[26] D.A. Powers, and Y. Xie, Statistical Methods for Categorical Data Analysis : 2nd edition, Emerald Group Publishing Limited, 2008.