Prediction of Major Features for Major Adverse Cardiac Events from KAMIR Dataset using Novel Nominal Gini-Index Algorithm

Heum Park

Department of Transportation Engineering, Youngsan University, Yangsan-si Kyungnam-do, Republic of Korea
hmpark@ysu.ac.kr

Abstract

The Korea Acute Myocardial Infarction Registry (KAMIR) dataset has been under construction at 41 Primary PCI Centers in Korea since November 2005. Many studies for the KAMIR have proceeded via analysis of statistical approaches: student's t-test, χ^2-test, and multivariate logistic regression analysis. However, there are problems, in that features tested are selected by domain experts according to the analysis conditions, that degrees of importance for features cannot be obtained, and that the huge numbers of features and instances involved incurs a high computation load and low processing speed. Thus, we considered novel feature selection methods using Gini-Index for prediction of the major features and reduction of feature space dimension. Unfortunately, only few studies on Gini-Index based nominal feature selection have as yet been completed, and problems in extracting representative features remain for 1) unbalanced dataset for classes, 2) instances having almost all of the features of the datasets, and 3) instances having almost all features with non-null values. Thus, for the datasets, the features selected are not discriminated for each class. In an effort to solve these problems and enable obtainment of good representative features for each class, we introduce here a novel Gini-Index feature selection algorithm for nominal datasets. We tested the algorithm for prediction of major features of AMI patients from the KAMIR. In the results, it can shows the degrees of importance for features with Gini values, and select the major features for given conditions without help by experts.

Keywords: *Gini-Index, Nominal Gini-Index, Feature Selection, Acute Myocardial Infarction, Korea Acute Myocardial Infarction Registry*

1. Introduction

Since November 2005, online registration of Korea Acute Myocardial Infarction patients (KAMIRs) has been carried out all of the 41 primary percutaneous coronary intervention (PCI) centers supported by the Korean Circulation Society (KCS), as reported in KCS's 50th anniversary memorandum. Many studies on predictions of major risk factors or treatment strategies for Korean Acute Myocardial Infarction (AMI) patients have been undertaken with respect to various clinical characteristics or treatment strategies [1].

The typical clinical characteristics include the influence of weather on daily hospital admissions for AMI [8], the impact of gender differences on long-term outcomes after PCI in patients with AMI [9], the obesity paradox in Korean patients undergoing primary PCI in acute ST segment elevation myocardial infarction (STEMI) [12], the predictors of six-month major adverse cardiac events (MACE) in 30-day survivors after AMI [14], a new risk score system for assessment of clinical outcomes in patients with non-STEMI [15], gender differences in success rates of PCI and short-term cardiac events [18], a hospital discharge risk score system for the assessment of clinical outcomes in patients with AMI [20], and others.

As regards treatment strategies, issues have included the impact of initial treatment delay on mortality among primary angioplasty patients with AMI [5]; triple versus dual antiplatelet therapy in patients with acute STEMI undergoing primary PCI [6]; the clinical safety of drug-eluting stents for AMI patients [7]; periodic variation and its effects on management and prognosis of Korean AMI patients [10]; comparison of outcomes between zotarolimus- and sirolimus-eluting stents in patients with STEMI [11]; evaluation of clinical outcomes and prognoses of the patients with near-normal coronary angiograms [13]; the safety and benefits of early elective PCI after successful thrombolytic therapy for AMI [16]; intensive pharmacologic treatment in patients with acute non-STEMI who did not undergo PCI [17], and current management of AMI for STEMI and non-STEMI patients [19], among others.

The typical statistical tools are SPSS and SAS, many studies having employed them for prediction, testing and verification purposes through analysis of student’s t-tests, χ^2-tests, and multivariate logistic regression analyses. The student’s t-test is applied to continuous variables and performed for categorical variables using the χ^2-test or Fisher’s exact test, and multivariate logistic regression analysis is performed to assess the relation between predictor variables. However, these methods come with some restrictions: 1) features and analysis conditions for testing are selected by domain experts; 2) degrees of importance for features cannot be obtained (only evaluated by verification with the significance probability p-value), and 3) the huge numbers of features and instances incur a heavy computation load, resulting in low processing speed.

We considered alternative machine-learning approaches as well as construction of novel feature selection methods using Gini-Index to predict major features from nominal datasets. There are many feature selection methods for nominal datasets: Information Gain, I-GI, Relief-Fn, χ^2, G-statistics, Mutual Information, Expected Cross Entropy, Weight of Evid, Odds Ratio, Relief, Decision-Tree Filter, Cross-Entropy Filter, Focus, Branch Bound, Beam Search, POE+ACC, LVF and LVW, among still others [21-25].

Nonetheless, there have been few studies on Gini-Index-based nominal feature selection, and most of them have focused on feature selection for text classification with numeric datasets. Also, using the Gini-Index entails certain representative-feature extraction problems regarding 1) unbalanced datasets for classes, 2) instances having almost all of the features of datasets, and 3) instances having almost all features with non-null values. Thus, for those datasets, the features selected by feature selection methods are not discriminated for each class. Specially, the KAMIR consists of the instances having almost all features (over 90%) and most features of the instances have non-null values, and that dataset is always unbalanced according to the conditions for analysis.

In the interests of solving these problems and enabling obtainment of good representative features from datasets, we suggest a novel Gini-Index feature selection algorithm for the nominal datasets, using evaluation by the feature pairs and their nominal values for each class, and by sum of them for the classes. In the following Section 2, we introduce both the existing Gini-Index algorithms and the Improved Gini-Index algorithm for text feature selection. In Section 3, we propose a novel Gini-Index algorithm for nominal datasets. In Section 4, we introduce the KAMIRs and the experimental dataset pertaining in the present study, along with the prediction results for the major KAMIR features. In Section 5, we draw conclusions and consider future work.

2. Existing Gini-Index for Text Feature Selection

In early work, the Gini-Index was used as a measure for determining the most appropriate splitting attribute at each node in a decision tree and for achieving, thereby, enhanced categorization precision. The more recent studies on the Gini-Index typically have concerned feature construction for genetic programming with decision classification (Mohammed et al. 2004), varieties of decision tree induction algorithms using splitting methods based on Gini-Index (Pang-Ning Tan *et al.*, 2006), and a fuzzy decision tree algorithm Gini-Index (B. Chandra et al. 2009), among still others [26-28]. For numeric datasets, there has been the Gini-Index for text classification and an adaptive Fuzzy kNN classifier based on the Gini-Index (Wenqian Shang *et al.*, 2007), as well as the I-GI algorithm (Heum *et al.*, 2011) [25, 29].

The main idea behind Gini-Index theory is as follows. Suppose S is a set of s samples, and that these samples have k different classes (C_i, $i=1,...,k$). According to the differences between classes, we can divide S into k subsets (S_i, $i=1,...,k$). Next, suppose S_i is a sample set belonging to class C_i, and that s_i is the sample number of sets S_i. Then the Gini-Index of set S is

$$Gini(S) = 1 - \sum_{i=1}^{k} p_i^2 \tag{1}$$

P_i is the probability, estimated with s_i/s, that any sample belongs to C_i. Gini(S)'s minimum is 0, and all of the members in the set belong to the same class, signaling that the maximum useful information can be obtained. When all of the samples in the set distribute equally for each class, Gini(S) is at its maximum, indicating that the minimum useful information can be obtained [25, 29]. For feature selection in text classification, W. Shang *et al.*, (2007) presented a novel Gini-Index algorithm based on Gini-Index theory for text feature selection, incorporating the following with a function, namely *Gini-A*:

$$Gini(W) = \sum_{i=1}^{k} P(W \mid C_i)^2 P(C_i \mid W)^2 \tag{2}$$

In this formula, if feature W appears in every document of class C_i, the maximum value, Gini value=1, can be obtained. When the documents distribute evenly where W appears, the minimum Gini value is obtained [13]. When the documents distribute evenly where W appears, the minimum Gini value is obtained [29]. However, the Gini-Index still shows feature selection bias in text classification: specifically, for unbalanced datasets having a huge number of features, the Gini values of low-frequency features are low overall, and for high-frequency features, the Gini values are always relatively high irrespective of the distribution of features among classes [25]. Thus, Heum *et al.*, (2011) presented an improved Gini-Index algorithm to correct the bias for unbalanced classes, and reformulated it to yield expression (2), by normalizing the probability $P(W|C_i)$ with the logarithm base 2, which reduces the range of $P(W|C_i)$ and produces unbiased Gini values, as follows:

$$IGini(W) = \sum_{i=1}^{k} \left| 1/\log_2 P(W \mid C_i)^2 \right| P(C_i \mid W)^2 \tag{3}$$

They obtained unbiased feature values, eliminated many irrelevant general features while retaining many specific features, and thereby could improve the overall classification performances when the local dimensionality reduction (DR) method was used [25]. To predict the major features for the nominal datasets that consists of unbalanced classes and the instances having almost all features with non-null values: the KAMIR, we considered the evaluation for the features by computations of Gini values with the feature pairs and their nominal values for each class, or sum of them for the classes.

3. Novel Gini-Index Algorithm for Nominal Dataset

As noted above, in feature selection for numerical datasets, the improved Gini-Index algorithm could solve the problems concerning unbalanced data and application of features to a dataset, and showed good performances. However, there have been few studies on Gini-Index-based nominal feature selection. And there remain problems regarding extraction of representative features for 1) unbalanced datasets, 2) instances having almost all of the features of datasets, and 3) instances having almost all features with non-null values. Thus, for those datasets, when the features of the instances have different nominal values for each class, we can obtain the representative features for each class. However, if most feature pairs and their nominal values show a low degree of differences among classes, we cannot obtain the discriminated features for each class. In order to solve those problems and obtain good representative features for each class, we suggest a novel Gini-Index feature selection algorithm for nominal datasets.

3.1. Reformulated Gini-Index Expressions

To solve those problems and select representative features for nominal datasets with the feature pairs and their nominal values, first, we reformulated expression (2) as expression (4), namely *NGini-A*, as follows:

$$NGini_A(F,V) = \sum_{i=1}^{k} \frac{m}{m_i} P((F=f,V=v)\,|\,C_i)^2 P(C_i\,|\,(F=f,V=v))^2 \qquad (4)$$

In this expression, *f* is a feature, *v* is the nominal feature value, C_i is the *i-th* class and *P* is the probability with a pair feature and its nominal value in a class. We added m/m_i to the expression for unbalanced the classes, where *m* is the number of all instances in the dataset and m_i is the number of all instances within a specific class C_i. Further, we reformulated the features as the feature pairs and their nominal values, and we computed the Gini values of the features with 1 or 0 according to whether or not the pairs exist in an instance. $P((F=f, V=v)|C_i)$ was applied to expression (4) to solve the unbalanced problems of the feature pairs and their nominal values for the classes. For example, the frequency of a feature pair *f* and nominal value *v: (f, v)* in a class c_1 is 10, and the frequency of a feature pair *f* and nominal value *v: (f, v)* in a class c_2 is 100, and there are different Gini values between the classes c_1 and c_2. Thus, in this expression, we focused on the number of the feature pairs and their nominal values for classes, as well as the unbalanced sizes of classes. Second, we reformulated expression (3) as expression (5), namely *NGini-B*, as follows:

$$NGini_B(F,V) = \sum_{i=1}^{k} \left|1/\log_2 P((F=f,V=v)\,|\,C_i)^2\right| P(C_i\,|\,(F=f,V=v))^2 \qquad (5)$$

In this expression, the frequencies of feature pairs and nominal values are normalized by the logarithm base 2, which reduces the range of $P((F=f, V=v)|C_i)$ and produces unbiased Gini values [25]. Third, we amended expression (2) to expression (6), namely *NGini-C*:

$$NGini_C(F,V) = \sum_{i=1}^{k} \frac{m}{m_i} P(C_i\,|\,(F=f,V=v))^2 \qquad (6)$$

In this expression, we added m/m_i, the ratios of the total number of instances to the total number of classes, to the basic Gini-Index expression, and we excluded $P((F=f, V=v)|C_i)$ for feature pairs and their nominal values given a class C_i. Thus, with this expression, we also focused on the number of feature pairs and their nominal values, along with the unbalanced sizes of classes.

3.2. Novel Nominal Gini-Index Algorithm for Feature Selection

With the features selected using those Gini-Index expressions, we can reduce the high dimensionality of the feature space. In applying representative features to datasets, there are two distinct ways of viewing DR, according to whether the task is performed locally (i.e., for each individual category) or globally. Local DR is that which chooses feature sets of terms for classification under each category in turn. This means that different subsets of document sets are used when working with different categories. Global DR is that which chooses feature subsets for classification under all categories. Commonly used global goodness estimators are the maximum and average (or sum) functions. However, neither of these two functions captures how a feature is distributed over different classes [30].

All functions of feature selection methods are specified “locally” to a specific class c_i ; in order to assess the value of a feature *f* in a “global,” class-independent sense, either the sum $Vsum(f)=\sum||C||V(f, c_i)$ or the maximum $Vsum(f)=max||C||V(f, c_i)$ of their class-specific values $V(f, c_i)$ usually are computed. According to the feature selection method, the method that shows the better performance, between the two, generally is adopted. Commonly used global goodness estimators are the maximum and average (or sum) functions. A well discriminated feature will have skewed distribution across the classes [25, 30]. However, neither of these two functions captures how a feature is distributed over different classes.

In the present study, we tested both global and local DRs to select the representative features using Gini-Index expressions. For the global DR, first, we calculated the Gini values $F(f, v, c_i)$ of the feature pairs and the nominal values for each class, using the Gini-Index expressions (4)~(6), and summed them for all classes, $F(f, v)=\sum_i F(f, v, c_i)$. Second, we ranked the feature pairs and the nominal values according to their Gini values. Third, we selected nine representative feature subsets $F_j(f, v)$ from $F(f, v)$, *j* being the feature subset reduced by 10%**j*, for 10%, 20% and so on up to 90% (the dimensionality of feature spaces was reduced by 10%**j* for each subset from $F(f, v)$). The feature subsets for Gini-Index expressions (4)~(6) were selected recursively.

As regards the policy of feature selection using local DR, first, we calculated the Gini values $F(f, v, c_i)$ of the feature pairs and the nominal values for each class, independently, using expressions (4)~(6). Second, we ranked the pairs for each class according to their Gini values. Third, we selected nine representative feature subsets $F_j(f, v, c_i)$ from $F(f, v, c_i)$, *j* being the feature subset reduced by 10%**j*, for 10%, 20% and so on up to 90% (the dimensionality of feature spaces was reduced by 10%**j* for each subset from $F(f, v, c_i)$. For expressions (4)~(6), the feature subsets were selected recursively.

Thus, we can select the independent feature subsets for each class by $F_j(f, v, c_i)$ locally, and select the feature subset by $F_j(f, v_i)$ globally, from the ordered features. All features pairs and nominal values can belong to multi-classes. The novel Gini-Index algorithm used according to those policies is as follows:

```
Input: vector spaces of training datasets with feature pairs, nominal values and class labels

Output: vector spaces of all datasets, Gini values and feature subsets for each Gini expression
For all Gini expressions (4)~(6)
```

```
  For each feature pair and nominal value $(f, v)$ do begin
    Calculate $Gini(f, v)$ using expressions (4)~(6) for Global DR
  For $i=1$ to $k$ do
    Calculate $Gini(f, v, c_i)$ using expressions (4)~(6) for Local DR
End
Obtain ordered feature sets $F(f, v)$
Obtain ordered feature sets $F(f, v, c)$
For $j=1$ to $9$ do begin
  For all feature pairs and nominal values $(f, v)$
    Select feature subsets $F_j(f, v)$ for upper $j*10\%$ from $F(f, v)$
    Apply features of $F_j(f, v)$ to vector spaces of all datasets
  End
  For all feature pairs and nominal values $(f, v, c)$
    Select feature subsets $F_j(f, v, c)$ for upper $j*10\%$ from $F(f, v, c)$
    Apply features of $F_j(f, v)$ to vector spaces of all datasets
  End
End
```

The process of the novel nominal Gini-Index algorithm for feature selection using new Gini-Index expressions is as shown in Figure 1.

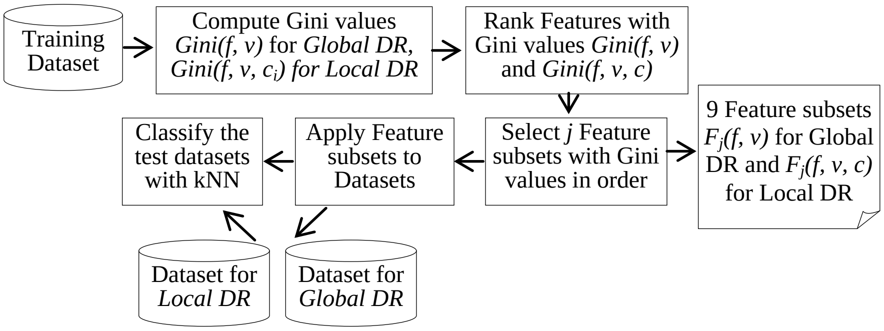
Training Dataset
Compute Gini values Gini(f, v) for Global DR, Gini(f, v, c_i) for Local DR
Rank Features with Gini values Gini(f, v) and Gini(f, v, c)
9 Feature subsets F_j(f, v) for Global DR and F_j(f, v, c) for Local DR
Select j Feature subsets with Gini values in order
Apply Feature subsets to Datasets
Classify the test datasets with kNN
Dataset for Local DR
Dataset for Global DR

Figure 1. Process of Nominal Gini-Index Algorithm to Obtain Feature Subsets for Global/local DRs

4. Experiments and Evaluation with KAMIR

4.1. KAMIRDataset

Since November 2005, online registration of Korea Acute Myocardial Infarction patients (KAMIRs) has been carried out at all of the 41 primary percutaneous coronary intervention (PCI) centers supported by the Korean Circulation Society (KCS), as reported in KCS's 50th anniversary memorandum. From the KAMIR dataset, we collected instances for 8,709 AMI patients who had experienced the Major Adverse Cardiovascular Events (MACE) between Nov 2005 and Dec 2010. We divided the instances into two classes, one having MACE within 12 months (1,856; 21.3% of patients) and the other not (6,853; 78.7% of patients); we then constructed the dataset with 167 features representing seven information groups, as shown in Table 1. The personal information of patients has five personal features, 13 for hospitalization, 39 for inspection, 21 for past medication, 64 for treatment, four for diagnosis and 21 for medication. The instances of the dataset have features almost all of which (over

90%) are of nominal values, the numbers of nominal values for each feature ranging from two to a maximum of 17. For experiments, we constructed a matrix from the above-noted data and instances according to the feature pairs and their nominal values (8,709 * 1019), and used 10-fold cross-validation for feature selection and evaluation.

Table 1. Seven Information Groups for 167 Features from KAMIR

Information		Features
Personal	5	Age, Gender, BMI, Central Obesity, Ratio of W/H
Hospitalization	13	First_medical_center, Vehicles, Transferred_from_other_hospital?, EKG_of_referred_hospital, initial_STE, initial_NSTE, initial_block, etc.
Inspection	39	Symptoms_on_admission, Chest_Pain, Dyspnea, Previous_angina_before_MI_symptoms, BPs, BPd, HR
Past Medication	21	ACEi, ARB, ACEi_or_ARB, Aspirin, BB, CCB, Cilostazol, Clopidogrel, DAPT, Digoxin, Diuretics, Eztrole, etc.
Treatment	64	Initial_therapeutic_strategy(plan),Initial_therapeutic_strategy_in_STEMI,Initial_therapeutic_strategy_in_NSTEMI, etc.
Diagnosis	4	Creatinine_on_admission, Total_cholesterol, HDL_cholesterol, LDL_cholesterol
Medication	21	Beta-blocker_in_hospital, Vastinan_in_hospital, Morphine_in_hospital, Statin_Fibrate_Vytorin_in_hospital, etc.

4.2. Experiments and Evaluations

To obtain the representative features, first, we selected the feature subsets using the new nominal algorithm according to the three new Gini-Index expressions (4)~(6) and with the global DR and local DR policies. Second, we applied the feature subsets to all of the datasets for each expression for those global DR and local DR policies. Third, we classified documents with the classification algorithm kNN. The kNN classifier has been widely used and offers good performance in various data classification areas. In the classification performance evaluations, we employed the F1 measure using Recall and Precision, where F1=(2*Recall*Precision)/(Recall +Precision). We compared the F1 measures classification performances for each feature selection according to the methods. For the purposes of the experiments, we developed the kNN classification tool, k being the number of classes for each dataset for kNN. We used 10-fold cross-validation for all of the classifications. Additionally, we compared the representative features for each method according to the results of the classification performances.

4.3. Experimental Results

First, we compared the classification performances for each new Gini-Index expression *NGini-A*, *NGini-B* and *NGini-C*, with the Local and Global DRs. Table 2 shows the classification performances Micro-F1 for all classes using the kNN classifier and for each of the upper 10% to 80% features, using the *NGini-A*, *NGini-B* and *NGini-C* with the Local and Global DRs. The performances for the Local DR were about 0.830 with the *NGini-A* and *NGini-C*, showed good results for the upper 30~50% features among the feature subsets. However, the *NGini-B* showed poor results, and did not indicate differences among the new Gini-Index expressions. Because the instances for the class having MACE within 12 months were 21.3% among all instances, and those of the other classes were 78.7%, these latter results (*NGini-B*) are meaningless for the purposes of this study. The performances for the Global DR were about 0.830 for all expressions, the *NGini-A* and *NGini-C* showing good performances for the upper 40~50% features among the feature subsets. The *NGini-B* showed good performances for the upper 10~20% and 70%~80% features.

Table 2. kNN Classification Performances of Micro-F1 for all Classes and each of Upper 10% to 80% Features using *NGini-A*, *NGini-B* and *NGini-C*, with Local and Global DRs

Features of Upper %	Micro-F1 for Local DR			Micro-F1 for Global DR		
	NGini-A	*NGini-B*	*NGini-C*	*NGini-A*	*NGini-B*	*NGini-C*
10%	0.787	0.787	0.215	0.818	**0.835**	0.389
20%	0.788	0.787	0.245	0.828	**0.834**	0.688
30%	0.810	0.787	**0.827**	0.824	0.820	0.824
40%	**0.832**	0.787	**0.832**	**0.830**	0.819	**0.839**
50%	**0.833**	0.787	0.790	**0.830**	0.820	**0.827**
60%	**0.831**	0.787	0.791	0.828	0.821	0.816
70%	0.828	0.788	0.815	0.827	0.835	0.812
80%	0.787	0.787	0.215	0.818	0.834	0.389

However, we focused on the features of the first class, because the goal of this study was to extract the major features of MACE within 12 months. Table 3 lists the kNN classification performances Micro-F1 for the class having MACE within 12 months and for each upper 10% to 80% features, using *NGini-A*, *NGini-B* and *NGini-C* with the Local and Global DRs. We also tested the basic Gini-Index replaced $P((F=f, V=v)|C_i)$ and $P(C_i|(F=f, V=v))$ instead of $P(W|C_i)$ and $P(C_i|W)$ from expression (2) (*Gini-A*). The *Gini-A* in Table 3 is the results of the basic Gini-Index. The performances for Local and Global DRs with the *NGini-C* (0.529 and 0.486) were better than those of the *Gini-A*, *NGini-A* and *NGini-B*. The performances for only the first class are low comparing with those of for all classes relatively, because the number of instances for the first class is 1,856 (21.3%) among all instances and almost instances have almost non-null values. However, we can see the performances of *NGini-C* are improved comparing with those of *Gini-A*, *NGini-A* and *NGini-B*.

Table 3. kNN Classification Performances Micro-F1 for Class having MACE within 12 Months and each of Upper 10% to 80% Features using *Gini-A*, *NGini-A*, *NGini-B* and *NGini-C*, with Local and Global DRs

Features	Micro-F1 for Local DR				Micro-F1 for Global DR			
Upper %	*Gini-A*	*NGini-A*	*NGini-B*	*NGini-C*	*Gini-A*	*NGini-A*	*NGini-B*	*NGini-C*
10%	0.000	0.000	0.000	0.393	0.370	0.319	0.417	**0.463**
20%	0.000	0.012	0.000	0.360	0.384	0.370	0.401	**0.464**
30%	0.046	0.206	0.000	**0.529**	0.336	0.324	0.301	**0.486**
40%	0.285	0.431	0.000	**0.450**	0.262	0.374	0.297	**0.463**
50%	0.424	0.423	0.000	0.030	0.332	0.373	0.315	0.353
60%	0.398	0.395	0.000	0.042	0.367	0.368	0.322	0.264
70%	0.380	0.364	0.005	0.255	0.370	0.359	0.417	0.243
80%	0.351	0.000	0.000	0.393	0.367	0.319	0.401	0.463

Table 4. Top Major Features, their Nominal Values and Gini Values for each Group with *NGini-C*

Information	Feature Pairs and their nominal values (Gini value)
Personal	Age under_65 (2.03), Gender female (1.52), BMI underweight (2.07).
Hospitalization	Sudden_Cardiac_Death yes (9.8), CPR Cardiopulmonary (mechanical) (7.6), CPR Cardioversion/defibrillation (5.8), Initial_ block yes (2.5), Initial_flat yes (2.5), etc.
Inspection	Heart_rhythm Flat_ECG (11.2), Heart_rhythm other_arrhythmia (9.5), Heart_ rhythm VT/Vfib (9.3), Heart_rhythm Sinus rhythm_ PVC (7.9), etc.
Past Medication	Eztrole yes (12.4), Digoxin yes (3.1), Nicorandil yes (2.6).
Treatment	Electrophysiology_study Planned (22), if_thrombolysis_is_contrain dicated Uncontrolled_hypertension (22), Treatment_after_failed_ PCI Death (22), etc.
Diagnosis	Creatinine_on_admission Abnormal (3.2)
Medication	anti-PLT_agents_-_add_warfarin no_anti_PLT_agent (13.9), anti-PLT_agents_ in_hospital no_anti_PLT_agent (13.9), anti-PLT_ag ents_code C000 (13.7), etc.

Second, among the representative features for *NGini-C*, we compared only the top major features of the first class (patients group having MACE within 12 months), for each of the information groups, Table 4 shows the features, their values and the Gini values over 2.0 for each information group. The first terms are the features for each information group, the second terms are their nominal values, and the values within round bracket are their Gini values by *NGini-C*.

5. Conclusions

The Gini-Index algorithm for nominal datasets has some problems in extracting representative features for 1) unbalanced dataset for classes, 2) instances having almost all of the features of the datasets, and 3) instances having almost all features with non-null values. Thus, we suggest a novel Gini-Index feature selection algorithm with three new expressions for nominal datasets. We adopted the Local and Global DRs to apply those features to the dataset, classified them, and compared their performances, for prediction of major features of AMI patients from the KAMIR. In experiments, we compared the classification performances Micro-F1 for all classes and for each of the upper 10% to 80% features, using the *NGini-A*, *NGini-B* and *NGini-C*, for the Local and Global DRs. For the class having MACE within 12 months, the performances of *NGini-C* for all DRs are better than those of *Gini-A*, *NGini-A* or *NGini-B*. Additionally, we presented the top major features with the Gini values using *NGini-C* for each information group. Thus, in the results, we can obtain the degrees of importance with the Gini values for features, and select the major features without domain experts if given conditions (class).

Acknowledgements

This research was supported by the Youngsan University research grant.

References

[1] K. H. Lee, M. H. Jeong, Y. K. Ahn and J. H. Kim, “Sex Differences of the Clinical Characteristics and Early Management in the Korea Acute Myocardial Infarction Registry” Korean Circulation Journal, vol. 37, **(2007)**, pp. 64-71.

[2] J. Y. Cho, M. H Jeong, Y. Ahn, S. C. Chae and I. H, Seong, “Predictive Factors of Major Adverse Cardiac Events and Clinical Outcomes of Acute Myocardial Infarction in Young Korean Patients”, Korean Circulation Journal, vol. 38, **(2008)**, pp. 161-169.

[3] S.Y. Hwang, “Comparison of Clinical Manifestations and Treatment-Seeking Behavior in Younger and Older Patients with First-time Acute Coronary Syndrome”, Journal of Korean Academy of Nursing, vol. 39, no. 6, **(2009)**, pp. 888-898.

[4] J. Y. Cho, M. H. Jeong, O. J. CHoi, S. Lee and S. Y, Jeong, “Predictive Factors after Percutaneous Coronary Intervention in Young Patients with Acute Myocardial Infarction”, Korean Circulation Journal, vol. 37, **(2007)**, pp. 373-379

[5] Y. B. Song, J. Y. Hahn, H. C. Gwon, J. H. Kim, S. H. Lee and M. H. Jeong, “The Impact of Initial Treatment Delay Using Primary Angioplasty on Mortality among Patients with Acute Myocardial Infarction: from the Korea Acute Myocardial Infarction Registry”, Journal of Korean Medical Science, vol. 23, **(2008)**, pp. 357-364.

[6] K. Y. Chen, S. W. Rha, Y. J. Li, K. L. Poddar and Z. Jin, “Triple Versus Dual Antiplatelet Therapy in Patients With Acute ST-Segment Elevation Myocardial Infarction Undergoing Primary Percutaneous Coronary Intervention”, American Heart Association Journals Circulation, vol. 119, **(2009)**, pp. 3207-3214.

[7] S. R. Lee, M. H. Jeong, Y. K. Ahn, S. C. Chae, S. H. Hur and Y. J. Kim, “Clinical Safety of Drug-Eluting Stents in the Korea Acute Myocardial Infarction Registry”, Circulation, vol. 72, **(2008)**, pp. 392-398.

[8] J. H. Lee, S. C. Chae, D. H. Yang, H. S. Park, Y. Cho, J. E. Jun and W. H. Park, “Influence of weather on daily hospital admissions for acute myocardial infarction (from the Korea Acute Myocardial Infarction Registry)”, International Journal of Cardiology, vol. 144, **(2010)**, no. 1, pp. 16-21.

[9] J. S. Woo, W. Kim, S. J. Ha, S. J. Kim, W. Y. Kang and M. H. Jeong, "Impact of gender differences on long-term outcomes after successful percutaneous coronary intervention in patients with acute myocardial infarction", International Journal of Cardiology, vol. 145, **(2010)**, no. 3.

[10] H. E. Park, B. K. Koo, W. Lee, Y. Cho, J. S. Park and J. Y. Choi, "Periodic Variation and Its Effect on Management and Prognosis of Korean Patients With Acute Myocardial Infarction", Circulation journal official journal of the Japanese Circulation Society, vol. 74, no. 5, **(2010)**, pp. 970-976.

[11] H. K. Kim, M. H. Jeong, Y. K. Ahn, J. H. Kim, S. C. Chae and Y. J. Kim, "Comparison of Outcomes Between Zotarolimus-and Sirolimus-Eluting Stents in Patients With ST-Segment Elevation Acute Myocardial Infarction", The American Journal of Cardiology, vol. 105, no. 6, **(2010)**, pp. 813-818.

[12] W. Y. Kang, M. H. Jeong, Y. K. Ahn, J. H. Kim, S. C. Chae, Y. J. Kim, S. H. Hur and Y. J. Kim, "Obesity paradox in Korean patients undergoing primary percutaneous coronary intervention in ST-segment elevation myocardial infarction", Journal of Cardiology, vol. 55, no. 1, **(2010)**, pp. 84-91.

[13] W. Y. Kang, M. H. Jeong, Y. K. Ahn, J. H. Kim, S. C. Chae, Y. J. Kim, S. H. Hur and I. W. Seong, "Are patients with angiographically near-normal coronary arteries who present as acute myocardial infarction actually safe", International Journal of Cardiology, vol. 146, no. 2, **(2011)**, pp. 207-212.

[14] J. H. Lee, H. S. Park, S. C. Chae, Y. Cho, D. H. Yang and M. H. Jeong, "Predictors of Six-Month Major Adverse Cardiac Events in 30-Day Survivors After Acute Myocardial Infarction (from the Korea Acute Myocardial Infarction Registry)", The American Journal of Cardiology, vol. 104, no. 2, **(2009)**, pp. 182-189.

[15] H. K. Kim, M. H. Jeong, Y. Ahn, J. H. Kim, S. C. Chae, Y. J. Kim, S. H. Hur, I. W. Seong and T. J. Hong, "A new risk score system for the assessment of clinical outcomes in patients with non-ST-segment elevation myocardial infarction", International Journal of Cardiology, vol. 145, no. 3, **(2010)**, pp. 450-454.

[16] D. S. Sim, M. H. Jeong, Y. Ahn, Y. J. Kim, S. C. Chae and T. J. Hong, "Safety and Benefit of Early Elective Percutaneous Coronary Intervention After Successful Thrombolytic Therapy for Acute Myocardial Infarction", The American Journal of Cardiology, vol. 103, no. 10, **(2009)**, pp. 1333-1338.

[17] H. C. Jeong, Y. K. Ahn, M. H. Jeong, S. C. CHae, J. H. Kim, I. W. Seong and J. Kim, "Intensive Pharmacologic Treatment in Patients With Acute Non ST-Segment Elevation Myocardial Infarction Who Did Not Undergo Percutaneous Coronary Intervention", Circulation Journal, vol. 72, no. 9, **(2008)**, pp. 1403-1409.

[18] K. H. Lee, M. H. Jeong, Y. K. Ahn, J. H. Kim, S. C. Chae and Y. J. Kim, "Gender differences of success rate of percutaneous coronary intervention and short term cardiac events in Korea Acute Myocardial Infarction Registry", International Journal of Cardiology, vol. 130, no. 2, **(2008)**, pp. 227-234.

[19] D. S. Sim, M. H. Jeong and J. C. Kang, "Current management of acute myocardial infarction: Experience from the Korea Acute Myocardial Infarction Registry", Journal of Cardiology, vol. 56, **(2010)**, pp. 1-7.

[20] H. K. Kim, M. H. Jeong, Y. Ahn, J. H. Kim, S. C. Chae, Y. J. Kim, S. H. Hur, I. W. Seong, T. J. Hong and D. H. Choi, "Hospital Discharge Risk Score System for the Assessment of Clinical Outcomes in Patients With Acute Myocardial Infarction (Korea Acute Myocardial Infarction Registry [KAMIR] Score)", The American Journal of Cardiology, vol. 107, no. 7, **(2011)**, pp. 965-971.

[21] M. Dash, H. Liu, "Feature Selection for Classification", Intelligent Data Analysis, vol. 1, no. 3, **(1997)**, pp. 131-156.

[22] Y. Yang and J. P. Pedersen, "A comparative study on feature selection in text categorization", Proceedings of the International Conference on Machine Learning, **(1997)**, pp. 412-420.

[23] D. Mladenic, "Feature subset selection in text-learning", Proceedings of the 10th European Conference on Machine Learning ECML98, **(1998)**.

[24] H. Park and H. C. Kwon, "Extended Relief-F Algorithm for Nominal Attribute Estimation in Small-Document Classification", IEICE Trans, vol. E92-D, **(2009)**, pp. 2360-2368.

[25] H. Park and H. C. Kwon, "Improved Gini-Index Algorithm to Correct Feature-Selection Bias in Text Classification", IEICE Trans, vol. E94-D, **(2011)**, pp. 855-865.

[26] M. A. Muharram and G. D. Smith, "Evolutionary Feature Construction Using Information Gain and Gini-Index", Lecture Notes in Computer Science, vol. 3003, **(2004)**, pp. 37-388.

[27] P. N. Tan, M. Steinbach and V. Kumar, "Introduction to Data Mining", Addison-Wesley, **(2006)**.

[28] B. Chandra and P. P. Varghese, "Fuzzifying Gini-Index based decision trees", Expert Systems with Applications, vol. 36, no. 4, **(2009)**, pp. 8549-8559.

[29] W. Shang, H. Huang, H. Zhu, Y. Lin, Y. Qu, and Z. Wang, "A Novel feature selection algorithm for text categorization", Expert System with Application, vol. 33, **(2007)**, pp. 1-5.

[30] F. Sebastiani, "Machine learning in automated text categorization", ACM Computing Surveys, vol. 34, no. 1, **(2002)**, pp. 1-47.