34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Identifying Relatively High-Risk Group of Coronary Artery Calcification Based on Progression Rate: Statistical and Machine Learning Methods

Ha-Young Kim, Sanghyun Yoo, Jihyun Lee, Hye Jin Kam, Kyoung-Gu Woo,
Yoon-Ho Choi, Jidong Sung and Mira Kang

Abstract— **Coronary artery calcification (CAC) score is an important predictor of coronary artery disease (CAD), which is the primary cause of death in advanced countries. Early prediction of high-risk of CAC based on progression rate enables people to prevent CAD from developing into severe symptoms and diseases. In this study, we developed various classifiers to identify patients in high risk of CAC using statistical and machine learning methods, and compared them with performance accuracy. For statistical approaches, linear regression based classifier and logistic regression model were developed. For machine learning approaches, we suggested three kinds of ensemble-based classifiers (best, top-*k*, and voting method) to deal with imbalanced distribution of our data set. Ensemble voting method outperformed all other methods including regression methods as AUC was 0.781.**

I. Introduction

The primary cause of death in advanced countries is coronary artery disease [1]. Coronary artery calcification, which is measured by computerized tomography (CT) scan, indicates the progression of atherosclerosis or accumulation of plaques in arteries, hence it has been known as a risk factor of CAD [2]. Thus, people whose CAC progresses faster than others should be considered at relatively high risk of CAD compared to the general population. Due to this reason, early prediction of progression rate of CAC makes possible to prevent CAD from developing into severe symptoms and diseases.

Many previous studies suggested particular risk factors that have statistical differences between disease and normal groups [6]. However, they could not provide the total impact of various risk factors. Although many test results, not limited to CAC related ones, are available in many hospitals, they are not fully utilized to predict the progression of CAC.

Also, there have been some studies on the prediction of incident risk of CAD such as angina pectoris, myocardial infarction, and ischemia stroke. However, they did not focus on the prediction of CAC progression [12] .

In this study, we considered more than 200 test results from the regular medical checkup data, and developed predictive models for CAC progression using statistical regression and machine learning methods. For statistical approaches, linear regression based classifier and logistic regression model were developed. For machine learning approaches, we suggested three kinds of ensemble-based models (best classifier-based, top-*k*-based, and voting model) to deal with imbalanced class distribution of our data set.

H. Y. Kim, S. H. Yoo, J. H. Lee, H. J. Kam and K. Y. Woo are with the Samsung Advanced Institute of Technology, Korea (phone: 82-31-280-9822; fax: 82-31-280-9086; e-mail: [hayoung7.kim, sam.yoo, jihyun.s.lee, hyejin.kam, kg.woo]@samsung.com).

Y. H. Choi, J. D. Sung and M. R. Kang are with Center for Health Promotion, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea (e-mail: [yh38.choi, jidong.sung, mira90.kang] @samsung.com).

II. Problem Definition

As explicated in the previous section, CAC has been considered as the important risk factor of CAD. According to Agatston's criteria, the grade of CAC is stratified with CAC score (CACS) as class 0 (CACS = 0, normal), class I (0 < CACS < 100, low-risk), class II (100 ≤ CACS < 400, mid-risk) and class III (CACS ≥ 400, high-risk).

Meanwhile, personal velocity of deterioration is different from one person to another. One can have CAC growing faster, which means that he should get more intensive treatment than others even in the same class. Due to such reasons, it is important to identify relatively high risk people of CAC based on their progression rate. Here, the high-risk person indicates the one whose CAC progression rate is higher than the average rate of the same group. To distinguish the high-risk people, the annual CAC progression rate (ACPR) for each subject is defined as follows:

$$ACPR = \frac{\ln(CACS_{follow-up}) - \ln(CACS_{base})}{year_{follow-up} - year_{base}} \quad (1)$$

where $CACS_{base}$ and $CACS_{follow\text{-}up}$ are positive. We exclude the one whose CACS is zero in this study. The relative risk of CAD for the one who has CAC is 10 times higher than that for the one who does not [7]. Therefore, the one whose CACS is greater than zero (i.e., class I, II, or III) should be cared through periodic follow-up observations.

It has been reported that depends on initial CACS and age, the CAC progression rates vary [1][3][5][6]. Thus, we performed ANOVA tests with three CAC classes and various age bands to make groups according to ACPRs that were statistically different from each other (we defined the follow-up period is 4-years). We confirmed that there are four groups based on the initial CAC class and age as follows: 1) CAC class I, 2) CAC class II and age ≤ 50, 3) CAC class II and age > 50, and 4) CAC class III. The average ACPR for each group was 18.87, 74.21, 59.97, and 128.32, respectively.

After that, we made binary classifiers to identify whether a person was at high-risk or not. We labeled the subject positive

if his ACPR was higher than the average ACPR of the group that he belongs to. Otherwise we labeled him as negative. Belonging to the positive class does not necessarily mean that he is at higher risk than all others in the negative class. It indicates his relative risk in the same group based on CAC class and the age group.

III. Data Set

This study was performed with a regular medical check-up dataset from the Samsung Medical Center in Seoul, Korea from 2003 to 2011. Since there were not enough data for women, we only selected records of men, who took at least two times of CAC CT scan during a 4-year interval. As a result, 580 men were considered in this study, and after labeling, the number of positive subjects were 203 and 377 were negative.

About 200 attributes for each subject were collected from the data set. The history of CAD (angina, myocardial infarction) and cerebrovascular disease (stroke, cerebral infarction) were collected by interviews. The experience of CAD-related medicine (such as aspirin and warfarin) and the history or current status of hypertension, hyperlipemia, diabetes mellitus, and smoking were also included. From physical data and laboratory investigations, we tried to use as many attributes as possible. However, attributes having many missing values or whose values are extremely skewed in a particular category often decrease the accuracy of the prediction. Thus, attributes with over 70% of missing or over 95% of single value were eliminated. Some numeric attributes whose normal ranges are well-known were discretized into two or three categories (i.e., Low, Normal, and High), and the transformed attributes were added into the attribute set. In addition, several compound attributes such as LDL/HDL, triglyceride/HDL, HOMA-IR (= (glucose x insulin)/405) and QUICKI (= 1/(log(insulin)+log(glucose)) were added. As a result, 125 numeric and 56 nominal attributes were remained. Finally, outliers were detected and eliminated from the samples by using histogram and box plot, and skewed numeric attributes such as TSH and CRP were transformed into a log-scale to have a normal distribution.

IV. Methods

We developed classifiers based on two approaches: statistics and machine learning.

A. Statistical Regression methods

Regressions have been widely used for prediction model and classification in medical domain. Linear and logistic regression approaches were used to predict relatively high or low risk compared to the same CAC class and age group.

Linear regression based model: First, we proposed a model developed by using multiple linear regression. This model is based on the assumption that CACS increases linearly [3]. It was built through the following three steps:

i. Predict follow-up CACS. Multiple linear regression was used including feature selection step; 11 variables out of 194 variables were selected.

ii. Calculate the ACPR of each patient with predicted follow-up CACS.

iii. Perform binary classification according to calculated ACPR.

Because step ii and iii are straightforward, here we describe only step i, the feature selection step. Due to the assumption that every feature in the regression model should be independent from each other, we first defined highly correlated group, called *HCG*, using Pearson's correlation coefficient ($r > 0.65$), a measure of the strength of the association between two variables. By using this, we could easily determine the representative variable from ones that were identified as correlated with each other. Second, by univariate analysis, we selected attributes set A_{ua} so that each attribute in this set has p-value<0.2. Third, we checked if there is an attribute whose VIF (Variance Inflation Factor, which quantifies the severity of multi-collinearity) was larger than 10. If so, an attribute *a* that has the largest VIF was selected. If the *HCG* containing *a* had more than one attributes in A_{us}, only one attribute considered as the most important factor among them was selected, and the rest of them were removed from A_{us}. Continue this test until all VIFs were less than 10. Finally, a regression model was built with remaining attributes in A_{ua}. As a result, the following equation was used to predict CAC follow-up scores:

$$\hat{y} = 0.883 + 0.615\,x_1 + 0.188\,x_2 + 0.159\,x_3 + 0.296\,x_4 - 0.098\,x_5 + 0.003\,x_6 + 0.018\,x_7 - 0.01\,x_8 + 0.006\,x_9 + 0.035\,x_{10} + 0.001\,x_{11}$$

where x_1: initial CACS (p-value<0.001), x_2: diabetes mellitus (p=0.01), x_3 : hypertension (p=0.01), x_4 : history of cerebrovascular disease (p=0.05), x_5 : experience of CAD-related medicine (p =0.15), x_6: systolic blood pressure (p=0.09), x_7 : carcinoembryonic antigen (p=0.29), x_8 : calcitonin (p =0.22), x_9: segmented neutrophil (p=0.07), x_{10}: body mass index (p <0.001), and x_{11}: lipoproteins (p=0.18).

Logistic regression model: Since the problem is the binary classification, we generated the second model using the logistic regression. We selected the features for this model through the method described above. Table I listed variables used in the logistic regression and their odds ratios.

TABLE I. ODDS RATIO FROM THE LOGISTIC REGRESSION MODEL

Variables	Mean±SD	OR (95% CI)	p-value
Age	53.9±7.5	0.98 (0.95 to 1.01)	0.183
Edema	32.4±1.1	1.46 (1.18 to 1.81)	0.001
Globulin	2.9±0.3	2.20 (1.14 to 4.29)	0.019
Body Mass Index	24.9±2.4	1.17 (1.07 to 1.29)	0.001
Lipoproteins	27.4±36.5	1.01 (1.00 to 1.02)	0.037
Calcitonin	3.9±3.0	0.94 (0.88 to 1.01)	0.101
CEA	2.1±1.5	1.23 (1.06 to 1.45)	0.012
TIBC	315.8±41.8	1.00 (0.99 to 1.00)	0.289
K	4.13±4.2	1.93 (0.91 to 4.13)	0.088
Phosphorus	3.42±3.2	1.90 (1.08 to 3.42)	0.028
Ln(Base CACS)	1.85±3.5	1.58 (1.36 to 1.85)	<0.001
Hypertension(0/1)	361(62%)/219(38%)	2.49 (1.61 to 3.88)	<0.001
Diabetes(0/1)	494(85%)/86(15%)	2.04 (1.12 to 3.73)	0.019

OR indicates Odds Ratio; CI, Confidence Interval; CEA, Carcinoembryonic antigen; TIBC, Total Iron-Binding Capacity

B. Data Mining Methods

Among various machine learning algorithms, we chose four classification methods that showed good performance in general: Decision tree [17], MultiBoost [18], LogitBoost [19], and Bagging [20]. Then, we selected attributes as features that made the accuracy of each classifier the best.

Because it is almost impossible to exhaustively search the optimal attribute set from more than 180 attributes, we used a heuristic feature selection approach as follows:

First, we define the candidate attribute set, which consists of two sets (S_w, S_f). S_w is a set of 23 clinically well-known attributes related to heart disease such as age, blood pressure, cholesterol, smoking, diabetes, hypertension, and etc. S_f is a set of 43 attributes whose information gain is positive. Second, we investigated each attribute in the set S_w to see if excluding the attribute makes the accuracy of the classifier improved. If so, the attribute was removed from the final attribute set. Third, we also added each attribute in the set S_f to the final attribute set if the attribute contributes to improve the classifier accuracy. We implemented these steps in a program that can generate the optimized final attribute set automatically. Although such heuristic approach might fall in local optimum problem, it is a useful method considering the time-efficiency.

Our dataset was imbalanced between classes, i.e., the negative set is 1.5 times larger than the positive set. Although it was not as serious as other highly imbalance data set, the prediction of models built by traditional classification methods could be dominated by the majority class (negative class), since they were developed on the assumption of balanced class distribution. As a result, true positive (TP) rate was poor (under 0.5). Since positive class means high risk people, TP rate is usually considered more importantly than true negative (TN) rate in the medical domain. Therefore, we needed to improve TP rate of our prediction model.

To achieve this, we adapted the under-sampling method (random sampling without replacement), which is popularly used to make the data set balanced. However, the under-sampled data set was unstable: the final attribute set was highly diverse according to the sample and the classification algorithm. Hence, we sampled *n* times with different random seeds and built a classifier for each sample with refining the attribute set to the sample and classification algorithm using the feature selection method described above. Then, we made the final classification models by using three methods as follows:

i. Select the classifier having the best accuracy among *n* classifiers as the final classifier. Fig. 1(a) depicts this method. We call the model built by this method the best classifier-based model.

ii. Extract top-*k* frequently used attributes by *n* generated classifiers. Then, make a classifier with the extracted attribute set. Fig. 1(b) illustrates this method. The model generated using this method is called the top-*k*-based model.

iii. Make a decision through voting by classifiers made from each sample as in Fig. 1(c). The model built by this method is called the voting model**.**

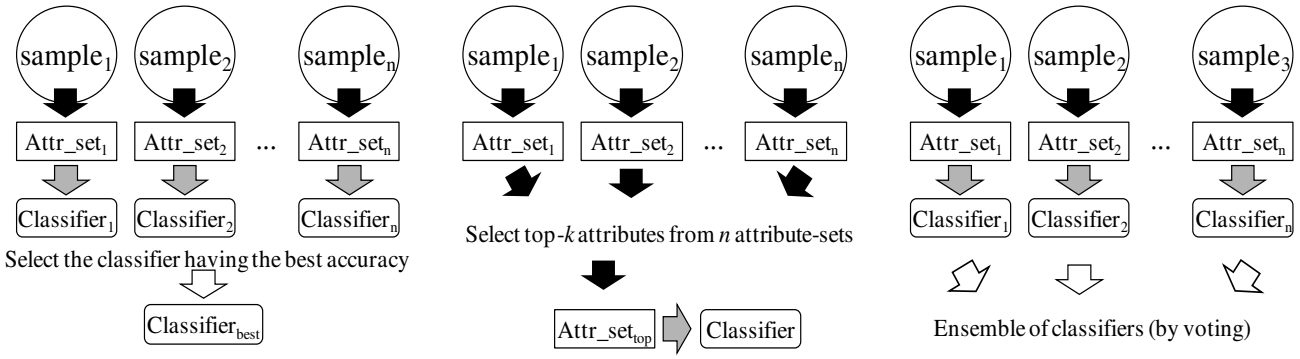
V. Experimental Results

We evaluated those five models described in the previous section by 5-folds cross-validation. All models except the linear regression based model were implemented by using Weka API [21]. In each fold, we performed the random sampling 10 times (i.e., n = 10) to make the training set balanced. All results are summarized in the Table II.

TABLE II. EXPERIMENTAL RESULTS BY 5-FOLDS CROSS-VALIDATION

Model	Correctly Classified	AUC	True Positive	True Negative
Linear R	64.5	-	35.9	94.8
Logistic R	68.0	74.9	67.5	68.5
Best	64.8	71.7	70.0	59.6
Top-5	68.2	73.1	73.4	63.1
Top-10	64.8	71.3	70.9	58.6
Top-20	63.8	70.0	67.0	60.6
Voting	71.4	78.1	73.9	69.0

With the constructed regression model (fitted R^2 = 0.766, predicted R^2 = 0.757, shrinkage = 0.9% by 5-fold cross-validation), the linear regression-based model correctly classified 64.5% of test instances. However, it seemed to under-estimate the risk as it had very high TN rate (94.8%) but poor TP rate (35.9%). The logistic regression model correctly



(a) Best classifier-based model (b) Top-*k*-based model (c) Voting model

Figure 1. Three classifier generation methods

classified 68.0% of test instances, and its TP and TN rate was 67.5% and 68.5%, respectively. The area under the ROC curve (AUC) was 74.9%. This result was better than other machine learning approaches except the voting model. For the best classifier-based model, LogitBoost was used as the classification algorithm. Its accuracy was lower (AUC =71.7%) than the logistic regression model. This was expected in fact, because the selected attributes were fit to only the sample from which they were found. It was reconfirmed that each sampled data set was unstable. Next, we tested top-*k*-based model. Although its accuracy was slightly better than that of the best classifier-based model (AUC=73.1%, *k*=5), the result was unsatisfactory compared to the logistic regression model. The rationale was that the best classifier-based method cannot exploit any of attributes that could contribute to classify instances correctly for some samples if they were not used frequently. However, we expect that this model could avoid the over-fitting problem. Also, the TP rate was better than that of the logistic regression method. On the other hand, the voting-based model had the best result as expected. Its AUC was 78.1%, and both of TP and TN rate were better than those of other models. The reason seems that the ensemble method can compensate the unstableness of samples by voting.

VI. Discussion and Conclusion

There have been many studies predicting CAD event with CACS as a predictor. However, to the best of our knowledge, there is no study about identifying high risk group of CAC. By predicting such a group, CAD events can be predicted and prevented in advance. For this purpose, we have suggested three kinds of ensemble-based classification methods to overcome imbalanced class distribution as well as two statistical regression models.

On the surface, the voting method seems to be better than statistical methods based on its accuracy. Despite of the low accuracy, however, statistical approaches have advantages such as their interpretability. Moreover, the linear regression-based model has the advantage that it can provide the estimated CACS in future, which may be used in further diagnosis or treatment. Meanwhile, machine learning approaches have the advantage that the features can be extracted automatically because it does not need any statistical analysis or assume any statistical condition. Therefore, we expect that both of two approaches can be used according to the purpose.

Because the data set used in this study was from one medical institution and only men were used to build suggested models, the data set is not representative of the public population. Thus, the accuracy of our models can be unstable. It means that applying these classifiers to women, other countries, and clinical institution, the accuracy could slightly decrease. One possible future work is to develop a model for women and more generalized model by collecting sufficient data from a variety of clinical institutions. Future study can include feature extraction methods instead of only selection and use of historical information to improve the accuracy of prediction models.

References

[1] Ambarish G, Khurram N, Sandy TL, Ferdinand RF, Lynn C, Matthew JB, Coronary calcium progression rates with a zero initial score by electron beam tomography: Int J Cardiol 177 227-231, 2007

[2] McClelland RL, Chung H, Detrano R, Post W, Kronmal RA, Distribution of coronary artery calcium by race, gender, and age; results from the Multi-Ethnic Study of Atherosclerosis: Circulation 113 30-37, 2006

[3] Richard AK, Robyn LM, Robert D, steven S, Joao Al et al., Risk factors for the progression of coronary artery calcification in asymptomatic subjects: results from the Multi-Ethnic Study of Atherosclerosis: Circulation 115 2722-2730, 2009

[4] Taylor AJ, Bindeman J, Feuerstein I, Cao F, Brazaitis M, O'Malley PG, Coronary calcium independently predicts incident premature coronary heart disease over measured cardiovascular risk factors: J Am Coll Cardiol 46 805, 2005

[5] James KM, Fay YL, David SG, Jonatha WW, Daniel SB, Leslee JS et al., Determinants of coronaty calcium conversion among patients with a normal coronary calcium scan: J Am Coll Cardiol 55(11) 1110-1117, 2010

[6] Yoon HC, Emerick AM, Hill JA, Gjertson DW, Goldin JG, Calcium begets calcium: progression of coronary artery calcification in asymptomatic subjects: Radiology. 24 236–241, 2002

[7] Budoff MJ, Lane KL, Bakhsheshi H, Mao S, Grassmann BO, Friedman BC et al., Rates of progression of coronary calcium by electron beam tomography: Am J Cardiol 86 8–11, 2000

[8] Callister TQ, Raggi P, Cooil B, Lippolis NJ, Russo DJ, Effect of HMG-CoA reductase inhibitors on coronary artery disease as assessed by electron-beam computed tomography: New Engl J Med 339 1972–1978, 1998

[9] Wong ND, Kawakubo M, LaBree L, Azen SP, Xiang M, Detrano R, Relation of coronary calcium progression and control of lipids according to National Cholesterol Education Program guidelines. Am J Cardiol 94 431–436, 2004

[10] Cassidy AE, Bielak LF, Zhou Y, Sheedy PF, Turner ST, Breen JF et al., Progression of subclinical coronary atherosclerosis: does obesity make a difference?: Circulation 111 1877–1882, 2005

[11] Bursztyn M, Motro M, Grossman E, Shemesh J, Accelerated coronary artery calcification in mildly reduced renal function of high-risk hypertensives: a 3-year prospective observation: J Hypertens 21 1953–1959, 2003

[12] Philip G, Laurie L, Stanley PA, Terence MD, Robert CD, Coronary artery calcium score combined with framingham score for risk prediction in asymtomatic individuals: J Am Med Assoc 291(2) 219-216, 2004

[13] Reinhard V, Paul C, Helmut S, Gerd A, Prediction of risk of coronary events in middle-aged men in the Prospective Cardiovascular Münster Study (PROCAM) using neural networks in middle-aged men in the PROCAM using Neural Network: Int J Epidermiol 31 1253-1262, 2002

[14] Matthew JB, Leslee JS, Sandy TL, Steven RW, Tristen PM, Philip FRF et al., Long-term prognosis associated with coronary calcification: J Am Coll Cardiol 49(18) 1860-1870, 2007

[15] Charles RT, Ann EN, Kevin BK, John M, Epidemiological data mining of cardiovascular bayesian networks: Electroinic Journal of Health Informatics 1(1) e3, 2006

[16] Tamar SP, Robyn LM, Neal WJ, Diane EB, Gregory LB, Alan DG et al., Coronary Artery Calcium Score and Risk Classification for Coronary Heart Disease Prediction: J Am Med Assoc 303(16) 1610-1616, 2010

[17] R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[18] G. I. Webb, MultiBoosting: A Technique for Combining Boosting and Wagging, Machine Learning, vol. 40, no. 2, 2000, pp. 159-196.

[19] J. Friedman, T. Hastie, R. Tibshirani, Additive Logistic Regression: a Statistical View of Boosting, Ann Stat, vol. 28, no.2, 2000, pp. 337-407.

[20] L. Breiman, Bagging predictors, Machine Learning, vol. 24, no. 2, 1996, pp. 123-140.

[21] R.R. Bouckaert, E. Frank, M.A. Hall, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, WEKA-experiences with a java open-source project, J. Machine leam. Res., vol. 11, pp. 2533-2541, 2010