35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Prediction of 4-year risk for Coronary Artery Calcification using Ensemble-based Classification

Jihyun Lee, Hye Jin Kam, Ha-Young Kim, Sanghyun Yoo, Kyoung-Gu Woo,
Yoon-Ho Choi, Jeong Euy Park and Soo Jin Cho

Abstract— **The progression of coronary artery calcification (CAC) has been regarded as an important risk factor of coronary artery disease (CAD), which is the biggest cause of death. Because CAC occurrence increases the risk of CAD by a factor of ten, the one whose coronary artery is calcified should pay more attention to the health management. However, performing the computerized tomography (CT) scan to check if coronary artery is calcified as a regular examination might be inefficient due to its high cost. Therefore, it is required to identify high risk persons who need regular follow-up checks of CAC or low risk ones who can avoid unnecessary CT scans. Due to this reason, we develop a 4-year prediction model for a new occurrence of CAC based on data collected by the regular health examination. We build the prediction model using ensemble-based methods to handle imbalanced dataset. Experimental results show that the developed prediction models provided a reasonable accuracy (AUC 75%), which is about 5% higher than the model built by the other imbalanced classification method.**

I. Introduction

As becoming an aging society, there is a rising interest on healthcare and wellness, i.e., living without illness. Among many chronic/acute diseases that people want to prevent, the coronary artery disease (CAD) is regarded as the most important one because it is the biggest cause of death nowadays. There are many guidelines to keep healthy and prevent CAD such as good nutrition, regular exercise, and a positive attitude (i.e., low stress), but the most important one is to assess one's current health status exactly. For this reason, many doctors recommend to get health examinations regularly. The purpose of the regular health examination is to screen for risk factors and diseases, evaluate health status, and provide preventive counseling interventions in an age-appropriate manner [1].

The coronary artery calcification (CAC) has been widely known as a risk factor highly related to CAD. CAC score (CACS) obtained by computerized tomography (CT) represents the amount of plaque accumulated in blood vessels and the degree of atherosclerosis. The growth of CACS has been regarded as a predictor of a future CAD. However, considering the cost of CT scanning, it is inefficient to perform as regular examinations if there is no risk. Thus, if people have no risk of CAC in the near future, whose coronary artery has not been calcified yet, they do not need to get CT scans again in a few years so that unnecessary medical costs can be reduced. Contrarily, if people are at high risk of CAC in the near future, they should be guided to get a CT scan again in a few years.

However, even for doctors, it is hard to determine the risk of CAC by regular health examinations, because dozens or even hundreds of direct/indirect risk factors related to CAC are measured per a person, which are from body measurements, lab tests, and medical interviews. Therefore, a tool such as a risk prediction model to assess the risk of CAC in general is needed. It takes many features such as age, BMI, blood pressure, blood glucose, cholesterol, smoking behavior and so on, and provides the probability that or the decision whether the coronary artery will be calcified in the near future (i.e., in 10 years). Because it is an easy way to summarize the influence of a lot of features on the CAC as a simple conclusion, it would be helpful to doctors and even to non-doctors.

There have been many studies about risk factors of the growth of CAC such as age, smoking, hypertension, diabetes, cholesterol and obesity [2][3][4]. However, most of them present only statistical differences between patients and non-patients with respect to each single risk factor. They do not address *n*-year prediction for a new occurrence of CAC by considering total effects of many risk factors. There is a prediction model for CAC progression [5], which is our previous work, but it can be applied to only high-risk groups, i.e., people to whom coronary artery calcification has already been started. Moreover, it focuses on CAC growing rate faster than average, not a new calcification event.

In this study, we develop a binary classification model identifying high risk people who are currently normal but likely to have a new CAC occurrence in the near future. According to [6], 62% of zero CAC Score (CACS) is lasted for 4-5 years, and the incidence of CAD increases at 0.1% per year in those cases. However, it has been reported that the risk increases by a factor of ten as CACS increases [7]. Thus, the occurrence of a detectable calcification (CACS > 0) within a medium term (i.e., 4-5 years) indicates that the risk of CAD becomes high. Thus, we focus on a new occurrence of CAC within 4 years.

To build the prediction model, we use a set of examination results that has been accumulated by the regular health examination. Due to the low prevalence rate, this dataset is imbalanced, which means that the number of people with CAC is much smaller than that of normal people; the normal group is 1.5 times larger than the abnormal group in our

J. H. Lee, H. J. Kam, H. Y. Kim, S. H. Yoo and K. G. Woo are with the Samsung Advanced Institute of Technology, Korea (phone: 82-31-280-9866; fax: 82-31-280-9086; e-mail:[jihyun.s.lee, hyejin.kam, hayoung7.kim, sam.yoo, kg.woo]@samsung.com).

Y. H. Choi, J. E. Park and S. J. Cho are with Samsung Medical Center, Korea (e-mail:[yh38.choi, jeongeuy.park, soojin77.cho]@samsung.com).

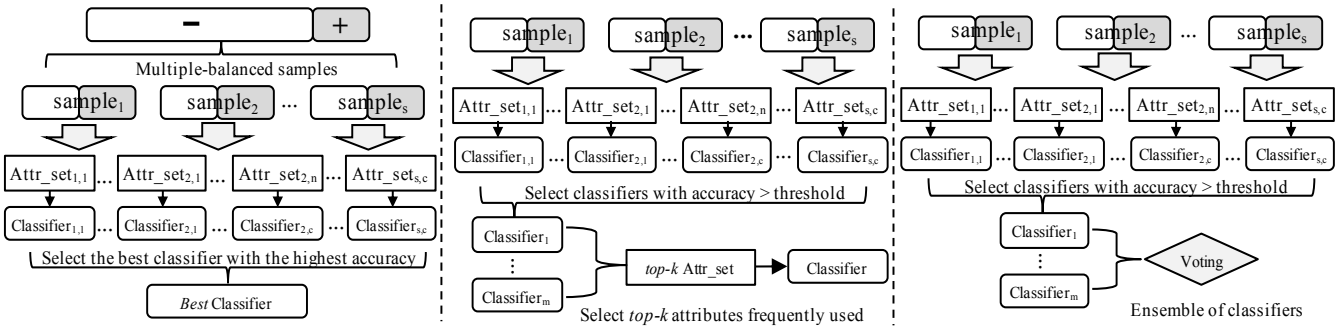

Figure 1. Three ensemble-based classification methods

dataset. Most of traditional classification algorithms assume that samples are evenly distributed among classes. Thus, the prediction models built by them tend to be dominated by the majority class (i.e., normal group) in the imbalanced dataset, thereby misclassifying a lot of instances from the minority class. For the effective classification on the imbalanced dataset, we adopt an under-sampling-based approach, which randomly removes samples of majority classes to make all classes balanced. In order to minimize the information loss due to biased under-sampling, we generate multiple classifiers based on multiple sub-sampling and integrate them through several ensemble strategies.

II. DATA SET

A. Data Collection and Class Definition

This study was performed with a regular medical check-up dataset from Samsung Medical Center from 2003 to 2011. Among the whole dataset , 836 men who took at least 2 times of coronary artery CT scan during a 4-year interval and had initial CACS = 0 were selected. Due to the insufficient size of the sample set to make an independent model, women were excluded in the modeling. We intended to classify people whose follow-up CACS in 4 years became non-zero into the high-risk (i.e., positive) group and the rest into the normal (i.e., negative) group. The size of positive and negative group is 288 and 548, respectively.

B. Preprocessing

We have considered about 200 features obtained along with the initial CACS at the same time, that were collected from various sources such as demographic data, physical measurements, lab tests, and interviews. The experience of CAD-related medicine such as aspirin and warfarin was considered. The history and current status of hypertension, hyperlipemia, diabetes mellitus, and smoking were also included.

From physical measurements and lab tests, we tried to use as many features as possible. However, since features having a lot of missing or extremely skewed values in a particular category often lead to inaccurate prediction model, those with over 70% missing or over 95% single value were eliminated. Some numeric features whose normal ranges are well-known were discretized into two or three categories (i.e., Low, Normal, and High), and then the transformed features were added into the feature set. In addition, several compound features such as LDL/HDL, triglyceride/HDL, QUICKI (= 1 / (log(insulin) + log(glucose)), HOMA-IR (= (glucose x insulin)/405) were added. As a result, 125 numeric features and 56 nominal features were extracted through the preprocessing step. By using histogram and box plot, outliers were detected and eliminated from the sample set. TSH and CRP were highly skewed, so they were transformed into the log-scale to have a normal distribution.

III. METHODS

A. Data-Driven Feature Selection

In the medical domain, most of studies on the development of prediction/classification models have considered several well-known outcome-related features. On the other hand, we decided to consider as many features as possible so that our model can utilize ones that are not well-known but can improve the prediction accuracy by interacting with other features.

The features were selected by the wrapper-based approach [8], which decides a feature set providing the highest accuracy for a given classification algorithm. Since it is impractical to exhaustively find the optimal feature set from several hundred features, we use a heuristic approach as follows: To speed up building a classification model and guarantee its reasonable accuracy, our feature selection method starts with the following well-known CAD-related features (predefined set): age, BMI, SBP, DBP, Total-cholesterol, Fasting glucose, BUN, Creatinine, CRP, Triglyceride(TG), HDL-cholesterol (HDL), LDL-cholesterol(LDL), TG/HDL, Hemoglobin A1c, Apolipoprotein A1 (ApoA1), Apolipoprotein B (ApoB), NT-proBNP, Lipoprotein (a), QUICKI, Microalbumin, Hypertension, Smoking, and Diabetes. First, it performs backward elimination from this predefined set toward increasing the classification accuracy. Then, our method completes a final feature set by forward selections adding features not in the predefined set. During this step, only features whose information gain is more than zero were considered for the efficiency. As a result, our feature selection method finds a near optimal feature set specialized to a particular classification algorithm and a dataset .

B. Ensemble-based Classification

In order to overcome the misclassification problem on imbalanced dataset, we use an ensemble strategy based on multiple balanced sub-sampling [9]. After performing multiple sub-samplings without replacement, multiple classifiers are generated over the various sub-samples, and the final classification model is built from the generated classifiers.

The classification accuracy depends on the classification algorithm and features used. Moreover, since even the same classification algorithm and features provides diverse performances according to application domains and datasets, it is hard to develop a robust classification model for every dataset. Thus, we open opportunities to diverse classification algorithms to build the most robust classifier. Given *s* sub-samples and *c* classification algorithms, we first generate a set of candidate classifiers for each sub-sample using all possible classification algorithms; s x c classifiers are generated in all. Then, we build a final classifier based on the following three ensemble strategies as shown in Fig 1. The details are as follows:

- *Best classifier-base method*: This method selects the best classifier with the highest accuracy out of all over the sub-samples as the final classifier.
- *Top-k-based method*: This method extracts top-*k* features frequently used in some qualified classifiers whose accuracy is over a given threshold, and then generates a final classifier with those features.
- *Voting-based method*: This method selects a classifier with the highest accuracy from each sample. Then, it lets all of them involve in making the final decision by voting.

IV. Experimental Evaluation

A. Prediction Model Construction

We implemented a classifier generation framework for 4-year prediction of a new CAC occurrence to automatically operate from the feature selection to building the final classifier using Weka[1] API. The number of samples was ten and the set of classification algorithms was {Decision Tree [10], LogitBoost [11], MultiBoostAB [12], Bagging [13]} for the ensemble classification. We evaluated three ensemble methods described in the previous section.

In order to show the effectiveness of the ensemble-based classification, we compared it with the cost-sensitive classification using diverse classification algorithms. The cost-sensitive method handles the imbalanced classification by assigning a higher penalty cost to misclassified instances from the minority class. The ratio of negative to positive samples was given to the positive group as the penalty of the misclassification. Then, we performed the feature selection in the same manner as the ensemble method.

B. Prediction Result

We performed 5-fold cross-validation for each prediction model and evaluated them using AUC. First, we investigated the accuracy of prediction models according to the classification algorithms and sample sets. TABLE I shows the variance of the accuracy of each classification algorithm over diverse sub-samples. Even in a single classification algorithm, its accuracy was diverse according to sub-samples. The difference between the minimum and the maximum AUC of each algorithm was 10-13%. TABLE II lists the accuracy and the feature set of classifiers built by LogitBoost for each sample set. The feature set used in the classifiers and their accuracy are various from sample to sample. The results of other classification algorithms were omitted due to the space limit, but they show similar tendencies.

TABLE I. THE PREDICTION ACCUARCY (AUC %) ACCORDING TO ALGORITHMS IN ENSEMBLE CLASSIFICATION

Algorithm	Min	Avg	Max	StdDev
Decision Tree	59.9	67.4	72.7	0.025
LogitBoost	68.1	74.5	79.3	0.024
MultiBoostAB	62.1	69.0	73.0	0.024
Bagging	61.4	67.3	71.9	0.026

The performances of the prediction models based on our multi-sample-based ensemble classification and cost-sensitive classification are listed in TABLE III. In general, ensemble-based classifiers outperformed cost-sensitive based classifiers using diverse classification algorithms. The classifiers built by the cost-sensitive method tended to be still dominated by the majority class. As a result, the sensitivity of every classifier was poor (around 30-50%), which is usually considered more importantly than the specificity to a screening tool. The best classifier-based method selected LogitBoost as the classification algorithm because it showed the best performance. Although it showed the highest accuracy (AUC 78.1%) as a local prediction model (shown in TABLE II), its accuracy as the global prediction model dropped considerably (AUC 74.07%). To prevent building such a biased prediction model to a particular sample, the top-*k*-based method (k = 15) selected classifiers from multiple samples with high accuracy (AUC > 60%) and retrieved *k* features that were frequently used in the selected classifiers. However, the sensitivity of the top-*k*-based model was slightly worse than the best classifier-based model. The reason is that several features, which contributed to improve the performance of the specific sample and classification algorithm, were ignored in the final prediction model. The voting-based method compensated the defects of best classifier-based and top-*k*-based methods by fully utilizing the classifiers specialized to each sample as well as covering diverse classification algorithms and samples. Consequently, the voting-based model was superior to other models in terms of accuracy as shown in TABLE III.

V. Discussion and Conclusion

We tried to fully utilize the results of various examinations as a feature set, not limit to well-known factors. Also, our method performed an extensive search to find a near optimal feature set for each sample set. As a result, albumin/creatinine ratio, fibrinogen, tPA, uric acid, hyperlipidemia, the medication history, and so on were additionally used in the prediction models, which are clinically related to CAC progression, even though they were not included in the predefined feature set. Consequently, our feature selection method led to the precise classification.

For the ensemble classification, we did not include representative machine learning algorithms: SVM and MLP since they usually took much time to build a single classifier even though the accuracy was inferior to other algorithms used in our study. Also, they are very sensitive to the condition of input data such as the number of features, feature type, value distribution, and value scale difference. In other words, additional preprocessing specialized to each algorithm (i.e.,

[1] http://www.cs.waikato.ac.nz/ml/weka/, Weka is a publicly available machine learning toolkit

TABLE II. THE ACCUARY AND FEATURE SET OF LOGITBOOST-BASED CLASSIFIER FOR EACH SAMPLE

Sample No	AUC (%)	Sensitivity (%)	Specificity (%)	# of Attrs	Attributes
1	73.5	64.9	72.3	23	Age, BUN, Abdominal fat, HBcAb, SBP, LDL/HDL, QUICKI, Bilirubin, Total cholesterol, ApoB, ApoA1, Microalbumin, tPA, Fibrinogen, Serum Iron, Medicine, HTN, …
2	74.8	65.3	74.6	27	Age, edema, Abdominal fat, HBcAb, SBP, DBP, Uric Acid, BUN/Creatinine ratio, Total cholesterol, Lp(a) Lipoproteins, ApoB, ApoA1, Microalbumin, PT, tPA, Hemoglobin A1c, …
3	75.0	66.1	73.5	24	Age, BUN, edema, BMI, DBP, LDL/HDL, QUICKI, Bilirubin , Total cholesterol, ESR, PAI-1, Albumin/Creatinine Ratio, ApoA1, Serum Iron , Medicine, Hypertension, Hyperlipidemia, …
4	74.7	67.1	70.1	12	Age, BMI, Albumin, Rheumatoid Factor , DBP, LDL/HDL, TG/HDL, QUICKI, tPA, Smoking, Hypertension, Diabetes
5	76.1	69.3	71.4	27	Age, edema, Abdominal fat, Albumin, Rheumatoid Factor, SBP, LDL/HDL, Glucose, ALT, Bilirubin, Total cholesterol, ESR, Albumin/Creatinine Ratio, ApoB, Microalbumin, tPA, …
6	73.8	61.4	73.6	7	Age ,DBP, TG/HDL, QUICKI, Lp(a) Lipoproteins, ApoB, Hypertension
7	75.7	64.5	75.8	23	Age, BMI, SBP, TG/HDL, ALP , Total cholesterol, Albumin/Creatinine Ratio, Hypertension, Lp(a) Lipoproteins, ApoA1, Floate, LDL, Triglyceride, JNC_V, Medicine, Homocysteine, …
8	78.1	69.3	74.9	21	Age, BUN, Edema, HBcAb, HBsAb, SBP, LDL/HDL, ALP, Total Bilirubin, Lp(a) Lipoproteins, Apolipoprotein A1, Osteocalcin, Microalbumin, Hemoglobin A1C, Hypertension, Diabetes,…
9	68.2	63.6	65.8	12	Age, BUN, BMI, SBP, LDL/HDL, TG/HDL, QUICKI, Glucose, Lp(a) Lipoproteins, Smoking, Hypertension, …
10	74.1	66.7	74.0	14	Age, BUN, BMI, SBP, DBP, LDL/HDL, TG/HDL, QUICKI, Apo B, TPA, Hemoglobin A1C, Serum Iron, Smoking, Hypertension, …

TABLE III. THE PREDICTION ACCURACY OF 4-YEAR CAC INCIDENCE ACCORDING TO CLASSIFICATION MODELS

Model	Cost-sensitive classification				Multi-sample-based ensemble classification		
	Decision Tree	LogitBoost	MultiBoostAB	Bagging	Best	Top-15	Voting
AUC (%)	61.17	70.45	64.97	64.36	70.94	70.20	74.07
Sensitivity (%)	44.44	59.03	31.25	38.19	65.62	63.89	66.32
Specificity (%)	67.36	69.44	85.76	82.29	63.89	68.06	69.44

normalization, generating dummy variables for categorical variables) is necessary to obtain a good performance from them. Hence, we excluded them from the classification algorithm pool.

Since the best classifier-based method cannot be free from the overfitting problem, the deviation of accuracy among samples tended to be large. On the other hand, the top-*k*-based and the voting-based methods were expected to attenuate the effect of overfitting. However, the accuracy of top-*k*-based model was not higher than that of the best classifier-based model. As mentioned before, the reason seems that features contributing to improve accuracy for specific samples and classification algorithms were removed. The voting-based model had good accuracy as expected. Since opinions from diverse classifiers built on different samples were considered, it could effectively reduce a bias toward a particular sample and classification algorithm. Generally, our ensemble-based classification method provided reasonable and better accuracy than another imbalanced classification method i.e., cost-sensitive approach, even without labor-intensive pre-processing and parameter tuning to make precise classifiers. Also, since it does not require comprehensive knowledge about classification algorithms and domain, it has advantages on the generalization and deployments.

REFERENCES

[1] Rachel A. Lee, Thomas N. R., Periodic Health Examination, Encyclopedia of Public Health, 2002, Encyclopedia.com. 29 Jan. 2013.

[2] K. M. James et al., Determinants of coronaty calcium conversion among patients with a normal coronary calcium scan, Journal of American Colledge of Cardiology 55(11), pp. 1110-1117, 2010

[3] H. C. Yoon et al., Calcium begets calcium: progression of coronary artery calcification in asymptomatic subjects, Radiology 24, pp. 236-241, 2002.

[4] M. J. Budoff et al., Rates of progression of coronary calcium by electron beam tomography: American Journal of Cardiology 86, pp.8-11, 2000.

[5] H. Y. Kim et al., Identifying relatively high-risk group of coronary artery calcification based on progression rate: Statistical and machine learning methods, Proc. 34th of the IEEE EMBC, pp. 2202-2205, 2012.

[6] J. H. Mieres et al., The role of non-invasive testing in the clinical evaluation of women with suspected coronary artery disease: American Heart Association Consensus Statement, Circulation 111 pp.68-696, 2005.

[7] Mj Budoff, Prognostic value of coronary artery calcification, Vascular Disease Prevention 2, pp. 2-20, 2005.

[8] Mark A. Hall, Geoffrey Holmes, Benchmarking attribute selection techniques for discrete class data mining, IEEE TKDE, 15(3), pp. 1437-1447, 2003.

[9] Xu-Ying Liu, Jianxin Wu, Zhi-Hua Zhou, Exploratory Undersampling for Class-Imbalance Learning, IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics 39(2), 2009

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

[11] J. Friedman, T. Hastie, R. Tibshirani, Additive Logistic Regression: a Statistical View of Boosting, Annals of Statistics 28(2), pp. 337-407, 2000.

[12] G. I. Webb, MultiBoosting: A Technique for combining boosting and wagging, Machine Learning 40(2), pp. 159-196, 2000.

[13] L. Breiman, Bagging predictors, Machine Learning 24(2), pp. 123-140, 1996