논문 2010-3-13

효율적인 의료진단을 위한 앙상블 분류 기법

Ensemble Classification Method for Efficient Medical Diagnostic

정용규*, 허고은**

Yong-Gyu Jung, Go-Eun Heo

요 약 의료 데이터 마이닝의 목적은 효율적인 알고리즘 및 기법을 통하여 각종 질병을 예측 분류하고 신뢰도를 높이는데 있다. 기존의 연구로 단일모델을 기반으로 하는 알고리즘이 존재하며 나아가 모델의 더 좋은 예측과 분류 정확도를 위하여 다중모델을 기반으로 하는 앙상블 기법을 적용한 연구도 진행되고 있다. 본 논문에서는 의료데이터의 보다 높은 예측의 신뢰도를 위하여 기존의 앙상블 기법에 사분위간 범위를 적용한 I-ENSEMBLE을 제안한다. 갑상선 기능 저하증 진단을 위한 데이터를 통해 실험 적용한 결과 앙상블의 대표적인 기법인 Bagging, Boosting, Stacking기법 모두 기존에 비해 현저하게 향상된 정확도를 나타내었다. 또한 기존 단일모델 기법과 비교하여 다중모델인 앙상블 기법에 사분위간 범위를 적용했을 때 더 뚜렷한 효과를 나타냄을 확인하였다.

Abstract The purpose of medical data mining for efficient algorithms and techniques throughout the various diseases is to increase the reliability of estimates to classify. Previous studies, an algorithm based on a single model, and even the existence of the model to better predict the classification accuracy of multi-model ensemble-based research techniques are being applied. In this paper, the higher the medical data to predict the reliability of the existing scope of the ensemble technique applied to the I-ENSEMBLE offers. Data for the diagnosis of hypothyroidism is the result of applying the experimental technique, a representative ensemble Bagging, Boosting, Stacking technique significantly improved accuracy compared to all existing, respectively. In addition, compared to traditional single-model techniques and ensemble techniques Multi modeling when applied to represent the effects were more pronounced.

Key Words : Ensemble-based, Bagging, Boosting, Stacking, I-ENSEMBLE

I. Introduction

Data mining of large amounts of data is unknown in advance to find useful information refers to a process of inferred knowledge. A very large field of application of data mining and marketing, and financial services companies, including currently active in all areas are being applied. In modern society, the human pursuit of the quality of life, particularly human disease is our task to be solved forever. Accordingly, the growing interest in medical research and health and medical data mining and efficient algorithms and techniques that automatically through the various diseases that are based on diagnosis.

Disease for efficient mining methods include traditional research and the relationship between the probability of first symptoms of the disease predicted to attain through the Bayesian method and the correlation between independent variables, calculated by assuming, a Naive Bayesian techniques[1]. In addition, DNA-related disease research, sports and include the

*종신회원, 을지대학교 의료IT마케팅학과
**정회원, 을지대학교 의료산업학부 의료전산학전공
접수일자 2010.06.05 수정일자 2010.06.17

study of genetic algorithm and SVM techniques, have been actively. Furthermore, the accuracy of the model for better prediction and classification have been proposed various methods. The most common way of various existing techniques to combine and mix combined model and the mixed models. Ensemble, where the combined technique method is called. Typical methods combining models and mixed models Bagging (Breiman, 1996), Boosting (Freund and Schapire, 1996), and Stacking (Wolpert, 1992), etc., and in recent Random Forests (Breiman, 2001) new combination techniques such as have been introduced. Bagging and Boosting techniques such as mixing or combining the biggest reason to use supervised learning techniques, especially in the conventional method using only one more than predicted and the classification accuracy is seen. Typically, bond and mixed a variety of techniques than the traditional single technique is known to improve performance.

In this paper, an efficient ensemble techniques to predict disease IQR (Interquartile range) applies. Among them are trying to deal with in this paper, a domain that hypothyroidism is a disease. Hypothyroidism is a failure, also known as thyroid disease, the amount of thyroid hormone in the body needs to produce thyroid. This is a state specific symptoms appear. Chronic thyroiditis (Hashimoto disease) or radioactive iodine treatment, thyroid removal can occur after surgery. Congenital or early onset cases, when an idiot (cretinism) is called as an adult because it's too late physical development of the child have the physique, and is an idiot or imbecile.

The existing ensemble technique, Bagging, Boosting, Stacking, when both apply Interquartile range , previously increased by more than 0.4% compared to the accuracy could be ascertained.

The composition of this paper starts with an introductory chapter, Chapter 2 in this paper provides an introduction to the existing ensemble methods, proposed in Chapter 3 of ensemble Interquartile range discusses a range of techniques applied. Chapter 4 Experimental techniques for applying the proposed Hypothyroidism Diagnosis Category hypothyroidism, and finally conclude.

II. Related research

2.1 Bagging

Bagging (bootstrap aggregating) is the method developed by Breiman as learning algorithm made multiple copies, each of these lessons are combined the results. Each data set created by bootstrapping is copied to the learning algorithm learning, then the majority voting and the results are determined by simple averaging method. Bagging is an effective way to represent performance, as small changes in data affect the results of the learning algorithm is a large, unstable if the learning algorithm is used.

표 1. Bagging 알고리즘
Table 1. Bagging algorithm

1: K= bootstrap sample number
2: **for** I = 1 to k **do**
3: operate N size bootstrap sample D_i
4: training basic classifier C_i to bootstrap sample D_i
5: **end for**
6: $C^*(x) = \text{argmax}_y \sum_i \delta(C_i(x) = y)$
if argmax = true
$\delta(\bullet) = 1$
else $\delta(\bullet) = 1$

2.2 Boosting (AdaBoost)

AdaBoost (Adaptive Boosting) of the training data set, the learning outcomes of the performance degradation to induce a large contribution to the weight given to certain patterns and after training, committee of the results is a step by step how to combine. The purpose of the model before misclassification Category Recurrence of the sample by increasing the probability of selection difficult to classify cases in the intensive Recurrence is intended to attack. Boosting algorithm, mainly used to combine with weak learning algorithm, but, AdaBoost on a decision tree or neural network

ensemble is applied to.

표 2. AdaBoost 알고리즘
Table 2. AdaBoost algorithm

1: w={w_j=1/N|j=1,2,...,N}
Initialize weights of all instances N
2: Let k=boosting round number
3: **for** I=1 to k **do**
4: Create Training set D_i depending on w from D(substitution)
5: Training C_i from D_i
6: Apply C_i to all instances of original training set D
7: $\epsilon_i = \frac{1}{N}[\sum_j w_j de[SC_i(x_j) \neq y_j)]$
{calculate error rates as Combined weights}
8: **if** $\epsilon_i > 0.5$ **then**
9: w = {$w_j = 1/N | j = 1,2,...,N$}
{reinitialize weights of all instances N
10: go to step 4
11: **end if**
12: $a_i = \frac{1}{2} ln \frac{1-\epsilon_i}{\epsilon_i}$
13: update weights of all instances for
$w_i^{(j+1)} = \frac{w_i^{(j)}}{Z_j} \times \begin{cases} exo^{-aj} & (C_j(x_i) = y_i) \\ \exp^{aj} & (C_j(x_i) \neq y_i) \end{cases}$
14: **end for**
15: $C^*(x) = \arg max_y \sum_{j=1}^{T} a_j \delta(C_j(x) = y)$

2.3 Stacking

Bagging and Boosting learning algorithm, applying the results of two days to decide by a majority vote against. the Stacking algorithm is a typical meta learning algorithm of the way a number of predictions based on the top step of the algorithm in the final lesson. Meta learning method to absorb the deviation between the predicted classification algorithm to improve the performance effects, but difficult to describe the final results and analysis are part of disadvantages. Stacking the selected classifier, rather than to merge a number of disparate strategies to measure the performance of the classification algorithm cross-validated standard method is known as a more flexible and sophisticated technique.

표 3. Stacking 알고리즘
Table 3. Stacking algorithm

1: **Input :**
training set L, algorithm pool A, categories set C, number of maximum iteration K
2: **Model Generation :**
For each of CV iteration
divide the dataset by number of CV folds (e.g.1:9)
for each of A algorithms:
apply the learning algorithm to training examplest
store the resulting model
Meta Data Generation :
for each instance in holdout set
calculate class probability for each category
generate meta data by combining class probability
3: **Meta Model Generation :**
apply learning algorithm to meta data
Rebuild base classifiers on the entire training data
4: **Classification :**
apply a new instance to base-level models and the meta-level model

III. Ensemble technique applied IQR

Data on the distribution of the measure of the change in measurement methods (measure of variability), or dispersion of the measure (measure of spread) is called. Range of measures against the spread of the irradiated material as a set of values (m) $\{x_1,...,x_m\}$ and some more about the property x, equation (1) is defined as.

$$range(x) = \max(x) - \min(x) = x_{(m)} - x_{(1)} \quad (1)$$

A set of values to estimate the spread of more robust data to measure changes in the statistic that represents one of the IQR of 50% of the total data amount of data scatter is measured. In other words, all data in order to clean up after the release of a four-way split in the middle portion of 50% was obtained for a range of expression (2) and 3 with a Interquartile range, the difference is called IQR.

$$IQR(x) = x_{75\%} - x_{25\%} \quad (2)$$

Therefore, by applying IQR medical data for effective predictive categories is to help create an additional class of variables. In this paper, and the Extreme Value to generate outlier. that means more to outlier points, especially in the medical field looking for a specific patient with unusual symptoms, test results can occur through outlier and the maximum value because the correct way should be interpreted and applied. to classify correctly the value of class through a range of Interquartile outlier patient data to generate value and Extreme Value Analysis on the importance of outliers in the existing ensemble methods, increasing accuracy is determined by adding. The proposed I-ENSEMBLE Equation (3) are.

$$I-ENSEMBLE = IQR + ENSEMBLE \quad (3)$$

Fig. 1 outlines the proposed structure for I-ENSEMBLE is showing. Interquartile hypothyroid data to calculate a range IQR outlier, and it generates and the Extreme Value variable is achieved through effective medical diagnosis.

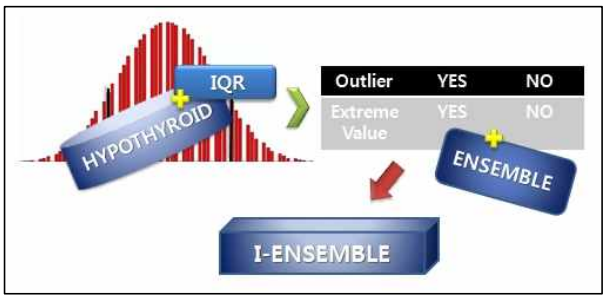

그림 1. I-ENSEMBLE 개념도
Fig. 1. Concept for I-ENSEMBLE

IV. Experimentation

4.1 Experimental methods

Hypothyroid data is applied in this paper. This data in order to determine whether the hypothyroidism was done 25 variables are included. 26th Class of the last variable determines whether the hypothyroidism as a result of a single data 3163 3012 Total private 95.2% 151 4.8%, the rest of the negative one represents the data with the data for the hypothyroid. If you apply this data IQR 27th and 28th variable outlier Extreme Value is created. What more outlier data points as a regular means that unlike a significant difference between the values shown. Extreme Value outlier is extremely similar to the meaning and value of the data shows the maximum number of means.

Table 4 shows the resulting two variables of the class. Outlier total of 3163 pieces of data that is classified as a class of 82 dogs showed 2.6%, Extreme Value include 86 dogs was 2.7%.

표 4. IQR 적용 후 데이터의 클래스
Table 4. After applying the data of the class, IQR

Class / Variables	Yes (%)	No (%)
Outlier	82(2.6)	3081(97.4)
Extreme Value	86(2.7)	3077(97.3)

IQR outlier data by applying the hypothyroid and Extreme Value is created by adding two variables. Created two variables to determine whether hypothyroidism is more of the ensemble algorithm acts as a variable, Bagging, Boosting, Stacking, when applying the results indicate more efficient than the existing. Existing tools were used to test the Weka Ensemble Algorithms Bagging, Boosting, Stacking was applied to the IQR.

4.2 Experimental results

Table 5 shows the experimental results of the existing ensemble of algorithms have been applied only when confirmed by comparing both the accuracy could be improved. Bagging the accuracy of, especially when applied to 100%, respectively, AdaBoost is 0.7964% of the cases increased, Stacking a whopping 2.055% of the cases could be confirmed or increased. The results in Fig. 2 can be found through the graph.

표 5. IQR 적용 후 정확도 변화
Table 5. The accuracy of the comparison of existing and IQR

Accuracy / Algorithms	Existing (%)	Apply IQR (%)
Bagging	99.2412	100
AdaBoost	99.0831 %	99.8735
Stacking	95.2261 %	97.2811

In addition, results when applied 2 times IQR and AdaBoost 100% accuracy of the technique, while the old, Stacking techniques applied in spite of the continuous 97.2811% did not change any more.

Table 6. Apply a single model, the accuracy of the comparison of existing and IQR

그림 2. 정확도 비교 그래프
Fig. 2. Comparison of the accuracy graph

Furthermore, the ensemble technique applied to the IQR to the other existing algorithms in order to prove more efficient than the Bayesian technique of existing algorithms. Bayesian technique results in accuracy was increased by 0.4426% 0.1879% rather than the Naive Bayesian techniques have decreased. In addition, rather than a single model applied to a single model to apply a comprehensive and multi-model ensemble techniques and thus a more efficient way by applying Interquartile range of ensemble techniques could indicate improved accuracy, apply a continuous range of Interquartile. When more may increase the accuracy was confirmed.

표 6. 단일 모델의 기존 및 IQR 적용 정확도 비교
Table 6. Apply a single model, the accuracy of the comparison of existing and IQR

Accuracy / Algorithms	Existing (%)	Apply IQR (%)
Bayesian	98.4825	98.9251
Naïive Bayesian	97.9134	97.7237

V. Conclusion

Effective diagnostic classification of medical data, the correlation between the variables considering the Bayesian learning, and to assume independence between variables, Naive Bayesian, DNA as a genetic algorithm and neural network research, SVM has been studied and various techniques. In addition to this model to better predict the classification accuracy of various methods have been proposed that the existing model of the ensemble method that combines a variety of techniques and methods typically Bagging, Boosting, Stacking exist.

In this paper, the higher the reliability of medical data to a range IQR ensemble technique was applied. What changes in data that represents the IQR as a measure and the three Interquartile ranges with Interquartile range of the data by specifying the calculation is to apply. Hypothyroidism or to classify the results of an experiment in medical data, ensemble technique, Bagging, Boosting, Stacking significantly improved accuracy compared to both existing and could be confirmed.

References

[1] P. Antal, et al., "Using literature and data to learn Bayesian networks as clinical models of ovarian tumors," Artificial Intelligence in Medicine, Vol 30, pp.257-281, 2004.

[2] Go-Eun Heo, Yong-Gyu Jung, Efficient Learning of Bayesian Networks using Entropy, The

institute Of Webcasting, Internet And Telecommunication, Vol 9, No 3, pp.31-36, 2009. 6

[3] Carvalho, D. R. and A. A. Freitas, Hybrid Decision Tree/Genetic Algorithm Method for Data Mining, Information Sciences, Vol 163, No 1-3, pp.13-35, 2004.

[4] Bauer, Eric and Ron Kohavi, An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting and Variants, Machine Learning Vol 36, pp.105-142, 1999.

[5] Zhang, Z. and C. Zhang, Agent-Based Hybrid Intelligent Systems, LNAI 2938, pp.127-142, 2004.

[6] Joon Hur, Jong Woo Kim, Characteristics on Inconsistency Pattern Modeling as Hybrid Data Mining Techniques, Journal Of Information Technology Applications & Management, Vol 15, No 1, pp.225-242, 2008.3.

[7] Conversano, C., R. Siciliano, and F. Mola, "Generalized Additive Multi-mixture Model for Data Mining", Computational Statistics and Data Analysis, Vol 38, No 4, pp.487-500, 2002.

[8] Breiman, L., Bagging Predictors, Machine Learning, Vol 24, pp.123-140, 1996.

[9] Brieman, L., "Random Forests", Machine Learning, Vol 45, No 1, pp.5-32, 2001.

[10] Breiman, L., Stacked Regressions, Machine Learning, Vol 24, pp.49-64, 1996.

[11] lan H.Witten and Eibe Frank, Data Mining, Addison Wesley, pp.315-333, 2005

[12] Kittler, J. et al., On combining classifiers, IEEE transactions on Pattern Analysis and Machine Intelligence, Vol 20, No 3, pp.226-239, 1998.

[13] Schaphire, Robert E., Theoretical views of boosting, In Computational Learning Theory: 4th European Conference EuroCOLT '99, 1999.

[14] Pang-Ning Tan & Michael Steinbach & Vipin Kumar, Introduction to Data Mining, ELSEVIER, pp.270-287, 2006

[15] Wolpert, L., Stacked Generalization, Neural Networks, Vol 5, No 2, pp. 241-259, 1992.

[16] Wolpert, D. and Macready, W., Combining Stacking with Bagging to Improve a Learning Algorithm, Technical Report, Santa Fe: Santa Fe Institute, 1996.

[17] Zhang, Z. and C. Zhang, "Agent-Based Hybrid Intelligent Systems", LNAI 2938, pp.127-142, 2004.

저자 소개

정 용 규(종신회원)

- 1981년 서울대학교 (이학사)
- 1994년 연세대학교 (공학석사)
- 2003년 경기대학교 (이학박사)
- 1999년~현재 을지대학교 교수
- 2001년~현재 ISO/TC154K위원장

<주관심분야: 임상데이터마이닝, 의료정보시스템, 전자거래표준>

허 고 은(정회원)

- 2007년~현재 을지대학교 의료산업학부 의료전산학전공

<주관심분야: 데이터마이닝, 비즈니스 인텔리젼스>