An Ensemble Method of Multiple Machines Learning System
แนววิธีแบบชุดของการทำงานระบบการเรียนรู้หลายระบบ

Souksavanh Choundaly (สุกสหวัน จุนดาลี)[1]* Dr.Wanida Kanarkard (ดร.วนิดา แก่นอากาศ)**

ABSTRACT

This research project presents two phases classification combined algorithmic approach to solve classification problem with heterogeneous classifiers. This novel approach is ensemble of C5.0, C&RT and CHAID models for creating and training of weak classifiers on difficult-to-classify patterns and combines the decision at 2-level. The approach is tested on six benchmark datasets from UCI machine learning repository. The results show that the proposed algorithm greatly outperforms existing methods, and achieves high accuracy.

บทคัดย่อ

โครงการวิจัยนี้ได้นำเสนอแนววิธีแบบชุดของการทำงานระบบการเรียนรู้หลายระบบแบบสองขั้นตอน เพื่อแก้ปัญหาการจำแนกแยกแยะด้วยเครื่องมือคัดแยกที่ต่างกัน แนวทางใหม่นี้เป็นการรวมชุดของ C5.0, C & RT และรูปแบบ CHAID สำหรับการสร้างและการฝึกอบรมของผลงานการคัดแยกกลุ่มที่อ่อนแอในชุดข้อมูลที่ยากต่อการจำแนกรูปแบบและผสมผสานการตัดสินใจที่เกิดขึ้นต่างระดับอยู่สองขั้นตอนโดยได้ทำทดสอบในหกชุดข้อมูลมาตรฐานจาก UCI ผลลัพธ์แสดงให้เห็นว่าแนววิธีที่นำเสนอมีความแม่นยำสูง สามารถทำงานได้มีประสิทธิภาพดีกว่าวิธีการเดิม

Key Words : Ensemble method, Classification
คำสำคัญ : แนววิธีแบบชุด การจำแนกข้อมูล

---

[1]*Correspondent author:choundaly@gmail.com*

** Student, Master of Computer Engineering Program in Computer Engineering, Faculty of Engineering, Khon Kaen University.*

*** Associate Professor, Department of Computer Engineering, Faculty of Engineering, Khon Kaen University.*

Introduction

One of the typical machine learning tasks is to create classification systems. Recently, in the area of machine learning there has been observed and increasing interest in combining multiple learning models in to one classification systems because single algorithms are specialized to solve some classes of learning problems, but cannot perform the best in all situation. One of the solutions to this limitation is to look for new methodologies allowing to an intelligence integration of different algorithms into one composed or hybrid system that can be used in diverse situations to improving prediction accuracy. Recent research in classification problems has mostly concentrated on ensemble methods [2, 4, 5] that construct a set of base classifiers instead of a single classifier. Both empirical observations and specific machine learning applications confirm that a machine learning algorithm outperforms all others for a specific problem or for a specific subset of the input data, but it is unusual to find a single expert achieving the best results on the overall problem domain. As a consequence multiple learner systems [9, 11, 14] try to exploit the local different behavior of the base learners to enhance the accuracy and the reliability of the overall inductive learning system. There are also hopes that if some learner fails, the overall system can recover the error.

Many machine learning algorithms apply local optimization techniques that may get stuck in local optima. For instance inductive decision trees employ a greedy local optimization approach, and neural networks [5] apply gradient descent techniques to minimize an error function over the training data. Building an ensemble using, for instance, different starting points may achieve a better approximation, even if no assurance of this is given.

Combining classifiers to achieve higher accuracy is an important research topic with different names such as combination of multiple classifiers, committee machine, classifier ensembles and classifier fusion. They are also, proposed to improve the classification performance of a single classifier. Interestingly, the recent works reviewed in Chapter 2 indicated that the ensemble of classifiers produced outstanding performances. Therefore, the main goal of this research is to propose a 2-level ensemble approach of well-known classification machine learning models to improve the performances of multiple classifiers.

Related works

Recent research on multiple classifiers, representing an interesting research direction to the development of highly accurate and specific classifiers. Several literatures are reviewed as follows: Luo and Cheng [12] introduced the ensemble predictor in breast mammography diagnosis. Chen and Tan [4] built the ensemble predictor for Type-2 diabetes. Chen, Zhang and Abraham [5] deployed the ensemble of neural networks for intrusion detection

system. Braha, Elovici and Last [2] applied the application of ensemble to the semiconductor control. Battiti and Colla [1]found that the use of a small number of neural nets (two to three) with a sufficiently small correlation in their mistakes reaches a combined performance that is significantly higher than the best obtainable from the individual neural networks.

Ho, Hull and Srihari [9] suggested a multiple classifier system based on rankings in the field of handwritten digit recognition. Nock [14], and Lam and Suen [11] analysed the application of majority voting to pattern recognition. Marques, Garcia and Sanches [13] analysed the behavior of the scoring mechanism in ensemble learning. Kuncheva [10] considered whether independence is good for combining classifiers. Their results support the intuition that ensemble of classifiers are better than independent classifiers. In fact there is a trade-off between accuracy and independence: more accurate are the base learners, less independent they are. Chawla, *et al.* [3] proposed a framework for building hundreds or thousands of classifiers on small subsets of data in a distributed environment. Evgeniou and Elisseeff [6] studied the ensembles of kernel machines such as SVMs. They found that the best SVM and the best ensembles had about the same test performance: with appropriate tuning of the parameters of the machines, combining SVMs does not lead to performance improvement compared to a single SVM. However, ensembles of kernel machines are more stable learning algorithms than the equivalent single kernel machine.

Hady and Schwenker [8] presented a cooperative semi-supervised learning approach for designing combining committee-based ensembles. Fumera and Roli [7] presented a theoretical and experimental analysis of linear combiners for classifier fusion. Their theoretical analysis shows how the performance of linear combiners depends on the performance of individual classifiers, and on the correlation between their outputs. Windeattand Ardeshir [18] constructed a decision tree for classifier ensembles.

These studies have proven the outstanding performances of ensemble classifier; therefore it is interesting to introduce the alternate method to improve the performance of multiple ensemble classifiers by a 2-level ensemble classifier as sometimes the double stages of cooperating work of human in making decision improve the decision results. In the next following chapters, experimental analysis of 2-level ensemble method of C5.0, C&RT and CHAID will be explained.

Theories

C5.0

C5.0 model is a variant and extension of a well-known decision tree algorithm, ID3 [15] and is a commercial version of C4.5 now widely used in many data mining packages such as Clementine and RuleQuest. It is targeted toward use with large datasets. The decision tree induction is close to that C4.5, but the rule generation is different. Unlike C4.5, the precise algorithms used for

C5.0 have not been revealed. C5.0 includes improvements to generate rules. Results show that C5.0 improves on memory usage by about 90 percent, runs between 5.7 and 240 times faster than C4.5 and produces more accurate rules. (need reference)

One major improvement to the accuracy of C5.0 is based on boosting. The splitting criterion of C5.0 algorithm is gain ratio that expresses the proportion of information generated by a split. The error-based pruning is used for C5.0's pruning. In the decision tree construction, a training dataset starts the root node of decision tree and a test is applied to search all possible splits for all predictor attributes to choose the best discriminates among the target variable and determine which child node will encounter next recursively. In the tree, one rule is created for each path from the root node to a leaf node in form of If–Then rule. Each attribute variable value along a path forms a conjunction in the rule antecedent. In addition, the leaf node determines the predicted class or value with the plurality rule, forming the rule consequent. The If–Then rules of decision tree are easier for human to understand. Even though C5.0 is relatively fast, building classifiers from large numbers of cases can take an inconveniently long time, especially when options such as boosting are employed.

C&RT

C&RT[18] is a form of decision tree that can be used for classification or regression problem. C&RT is a nonparametric technique used to explore the relationships between a target outcome variable and a large number of potential predictors. In C&RT settings, the first setting is the prior probability of the target variables (frequency of the classes). Another important setting is to select the measure of "data impurity" to use in evaluating variable split criteria. The Gini score is used to measure this impurity. The Gini score is based on the relative frequency of sub ranges in the predictor variables. It examines each predictor variable to identify the most significant predictor at each step to split the sample into two mutually exclusive and homogenous subgroups. Each C&RT split is an optimal balance between sensitivity and specificity for predicting the outcome variable. This process is conducted repeatedly until the sample is split into completely homogeneous groups or until a predetermined maximum level of splits is reached. The final result is a classification tree. The entire sample is referred to as the root, each split is referred to as a branch, and the data subset resulting from the split is called a node; the terminal or ending nodes are referred to as leaves.

CHAID

CHAID stands for Chi-squared Automatic Interaction Detector [18]. This technique constructs non-binary trees, for classification problems when the dependent variable is categorical relies on the Chi-square test to determine the best next split at each step. CHAID uses a Chi-square splitting criterion. More specifically, it uses the p-value of the Chi-square test. Specifically, the algorithm proceeds as follows:

1. Preparing predictors. The first step is to create categorical predictors out of any continuous predictors by dividing the respective continuous distributions into a number of categories with an approximately equal number of instants. For categorical predictors, the classes are defined.

2. Merging classes. The next step is to cycle through the predictors to determine for each predictor the pair of predictor categories that is least significantly different with respect to the dependent variable. For classification problems, it will compute a Chi-square test and F tests if the problem is the regression problems.

3. Selecting the split variable. The next step is to choose the split the predictor variable with the smallest adjusted p-value, i.e., the predictor variable that will yield the most significant split; if the smallest adjusted p-value for any predictor is greater than some alpha-to-split value, then no further splits will be performed, and the respective node is a terminal node.

4. Continue this process until no further splits can be performed (given the alpha-to-merge and alpha-to-split values).

Ensemble Model

Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a vote of their predictions. The benefit of ensemble machine learning is to combine a number of rough classifications into a more accurate aggregate class prediction.

In general, the ensemble method of the machine learning models can be divided into two categories: parallel {bagging} and serial approaches {boosting}.

Bagging [3] a name derived from bootstrap aggregation was the first effective method of ensemble learning and is one of the simplest methods of reusing data in order to improve classification. The method uses multiple versions of a training set by using the bootstrap, i.e. sampling with replacement. Each of these data sets is used to train a different model. The outputs of the models are combined by voting to create a single output. Bagging is only effective when using unstable (i.e. a small change in the trainingset can cause a significant change in the model) non-linear models.

In parallel approach, classifiers are trained in parallel using the same dataset. These independently trained classifiers are then combined using rules such as averaging voting, multiplying linear combination and other methods.

Boosting is an algorithm which can be viewed as a model averaging method. It is the most widely used ensemble method and one of the most powerful learning ideas introduced recently. Originally designed for classification, it can also be profitably extended to regression. One first creates a 'weak' classifier, that is, it suffices that its accuracy on the training set is only slightly better than random guessing. A succession of models are built iteratively, each one being trained on a data set in which points misclassified by the previous model are given more weight. Finally, all of the successive models are weighted according to their success and

then the outputs are combined using voting (for classification) or averaging (for regression), thus creating a final model.

Sohn and Lee [16] constructed by increasing the number of classifiers one-at-a-time to improve the classification accuracy. The dataset used to train each classifier is chosen based on the performance of the earlier classifiers in the series. The goal is to produce new classifiers that can somehow compensate for the weakness of the existing classifiers. A well-known example of the serial approach is boosting [6], which tries to generate new classifiers that are better able to correctly classify samples for which the current committee performance is poor.

Research design

The tested datasets are real-world problems obtained from the UCI repository of Machine Learning Databases [17]. Evaluation of classification accuracy is set to be the evaluation tool as well as the execution time in second of the test model. The design research process of the proposed ensemble method of machine learning models consists of two levels as show in Figure 5.

First Level Committee:

1. To analysis the ensemble of these 3 decision tree classifiers with 6 selected data sets which varied in term of class size, attributes and instances.

2. To reduce the chance of training model finding local minima. To ensure consistent results, the same random number initialization will be used for all model training for all experiments.

3. To characterize the generalization performances of the proposed ensemble methods, the accuracy reported is the average of the classification accuracy.

4. After training, several independently train machine learning models are aggregate in an appropriate linear combination manner. The linear combination method, as a linear combination of several machine learning models, includes: The majority voting, **Highest confidence wins** and Confidence-weighted voting.

Second Level Committee:

5. Purpose the 2-levels ensemble of C5.0+C&RT+CHAID to improve the performances of Test Case 2 by introducing the similar ensemble model to rework on the misclassification data.

Data sets

6 data sets from the UCI repository are selected to test in this research with no missing data. These data represent a wide range of domains and data characteristics. A brief description of the properties of the data is presented in Table 1. Some data consist of only symbolic, or a mixtures of symbolic and numeric (integer, floating point or both). The datasets store pre-specified training data derived from a variety of sources: medical, biological, or sociological domains.

Results

In this experiment, the 2-levels ensemble of 3 classifiers is presented. The question of this work is that adding more level of ensemble classifiers will improve the traditional single level ensemble or not. As many human tasks, performing by having two stages of committee could improve the decision results. However, this may not be true in many cases that do require the quick decision, as a result of having more time to work and complicated stages of adding more tasks. The 2-levels ensemble is introduced to work on the misclassification performed by the first level ensemble.

Table 2 provided the results confirmed that even though adding more tasks in making decision, this approach worth the improvement in term of accuracy. Clearly results could be seen in steel and bank data sets. For all 6 data sets, the misclassification could be fixed by adding another level of ensemble classifiers. The better performances are gained reasonably well regarding the extra execution time.

Figures 6 and 7 clearly display the big improvement in adding another level of ensemble to the classification task, although the overall execution time seems to be large. Additionally more second in all 6 data sets, but the extra time spent in the additional level of work still spent less than 1/3 of the original time. The outstanding performances of the proposed model could be found in big data set as Bank and Wine.

Conclusion

Machine concept learning has been a challenging research area since the birth of modern computers. Yet, a human apprentice is able to learn fairly complex concepts relatively easily by observing how competent experts make their classifications. Imitating the human behavior by introducing the ensemble machine learning has been proven to perform well in many real world problems. Complicated tasks, for example, in order to diagnose a disease, a diagnostician performs a sequence of tests. The choice of the next test to perform depends on the results from the earlier tests. This sequence provides valuable information as to which attributes are relevant to the expert in determining the classification, and how some of the attribute values affect the relevance of the other attributes. As a result, the learning process becomes more tractable. The problem of uncovering the outcomes done by the first party could be reviewed again in another stage of work by the common group of party. Therefore, this research investigates the new two-level ensemble of classifiers in the classification tasks. By nature, adding extra tasks will results in extra time and complication in making final decision. However, some classification tasks that do require the high accuracy without the constraint in time could benefit from this approach.

Ensemble methods are learning algorithms that construct a set of classifier and then classify new data points by

taking a weighted vote of their prediction. This thesis reviews these methods and explains why ensembles can often perform better than any single classifier. Some comparing ensemble methods are investigated and new experiments are presented to yield interesting new learning algorithms. The findings from this study have indicated that the proposed model could work well in big data sets and the extra time of adding another level of ensemble worth the extra improvement in classification accuracy. The choice of suitable number of classifiers also tested in order to gain the optimal performances comparing to the single classifier.

In summary, the proposed 2-levels ensemble of C5.0 + C&RT+ CHAID models yields the high accuracy in classification and could be generalized applied to any class size of data.

References

1. Battiti R, Colla AM. Democracy in Neural Nets: Voting Schemes for Classification. Neural Networks, 1994; 7(4): 691–707.
2. Braha D, Elovici Y, Last M. Theory of actionable data mining with application to semiconductor manufacturing control. International Journal of Production Research, 2007; 45(13): 3059-3084.
3. Chawla NV. *et al.* Learning Ensembles from Bites: A Scalable and Accurate Approach. Journal of Machine Learning Research. 2004; 5: 421–451.
4. Chen H, Tan C. Prediction of Type-2 Diabetes Based on Several Element Levels in Blood and Chemometrics. Biological Trace Element Research, 2012; 147(1-3): 67-74.
5. Chen YH, Zhang Y, Abraham A. Estimation of distribution algorithm for optimization of neural networks for intrusion detection system. Artificial Intelligence and Soft Computing - Icaisc 2006, Proceedings, 4029, 2006; 9-18.
6. Evgeniou TMP, Elisseeff A. Leave One Out Error, Stability, and Generalization of Voting Combinations of Classifiers. Machine Learning, 2004; 55(1): 71–97.
7. Fumera G, Roli F. A Theoretical and Experimental Analysis of Linear Combiners for Multiple Classifier Systems. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005; 27(6): 942–956.
8. Hady MFA, Schwenker F. Combining Committee-Based Semi-Supervised Learning and Active Learning. Journal of Computer Science and Technology, 2010; 25(4): 681-698.
9. Ho TK, Hull JJ, Srihari SN. Decision Combination in Multiple Classifier Systems. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1994; 16(1): 66–75.
10. Kuncheva LI, Lakhmi CJ. Designing Classifier Fusion Systems by Genetic Algorithms. IEEE Transactions on Evolutionary Computation, 2000; 4(4): 327–336.

11. Lam L, Ching YS. Application of Majority Voting to Pattern Recognition: An Analysis of Its Behavior and Performance. IEEE Transactions on Systems, Man and Cybernetics, Part A, 1997; 27(5): 553–568.
12. Luo ST, Cheng BW. Diagnosing Breast Masses in Digital Mammography Using Feature Selection and Ensemble Methods. Journal of Medical Systems, 2012; 36(2): 569-577.
13. Marques AI, Garcia V, Sanchez JS. Exploring the behaviour of base classifiers in credit scoring ensembles. Expert Systems with Applications, 2012; 39(11): 10244-10250.
14. Nock R. Inducing interpretable voting classifiers without trading accuracy for simplicity: Theoretical results, approximation algorithms, and experiments. Journal of Artificial Intelligence Research, 2002; 17: 137-170.
15. Quinlan JR. C4.5: Programs for Machine Learning, Morgan Kaufmann. 1993; [n.p.].
16. Sohn SY, Lee SH. Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea. Safety Science, 2003; 41(1): 1-14.
17. The UCI KDD Archive, Department of Information and Computer Science. Irvine: California. 1999.
18. Windeatt T, Ardeshir G. Decision tree simplification for classifier ensembles. International Journal of Pattern Recognition and Artificial Intelligence, 2004; 18(5): 749-776.

Table 1 Summary description of data sets

Data set	#Ex.	#Atts.	#Cl.
1. Wine	6,497	11	7
2. Steel plate	1,941	27	7
3. Car Evaluation	1,728	6	4
4. Yeast	1,484	8	10
5. CTG	2,126	22	3
6. Bank	45,211	16	2

Notes: (i) #Ex.: number of examples, (ii) #Atts.: number of attributes, (iii) #Cl.: number of classes.

Table 2 Evaluation of 2-levelsensemble of C5.0+C&RT+CHAID on 6 data sets

	1st-level Committee			2nd-level Committee			Overall Performances		
Datasets	**Correct**	**Wrong**	**Total**	**Correct**	**Wrong**	**Total**	**Correct**	**Wrong**	**Total**
Wine	5,819	678	6,497	664	14	678	6,483	14	6,497
	89.56	10.44		**97.94**	2.06		**99.78**	0.22	
Steel	1,940	1	1,941	1	0	1	1,941	0	1,941
	99.95	0.05		**100.00**	0		**100.00**	0.00	
Car	1,672	56	1,728	55	1	56	1,727	1	1,728
	96.76	3.24		**98.21**	1.79		**99.94**	0.06	
Yeast	1,212	272	1,484	170	102	272	1,382	102	1,484
	81.67	18.33		**62.50**	37.50		**93.13**	6.87	
CTG	2,089	37	2,126	36	1	37	2,125	1	2,126
	98.26	1.74		**97.30**	2.70		**99.95**	0.05	
Bank	42,061	3,150	45,211	3,150	0	3,150	45,211	0	45,211
	93.03	6.97		**100.00**	0		**100.00**	0.00	

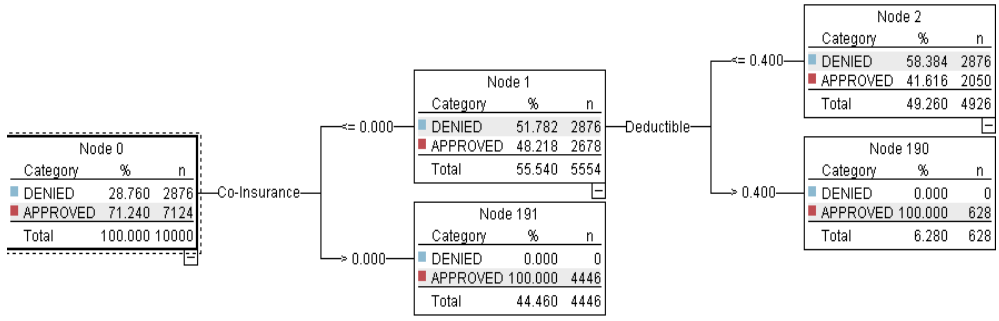


Figure 1 C5.0 tree

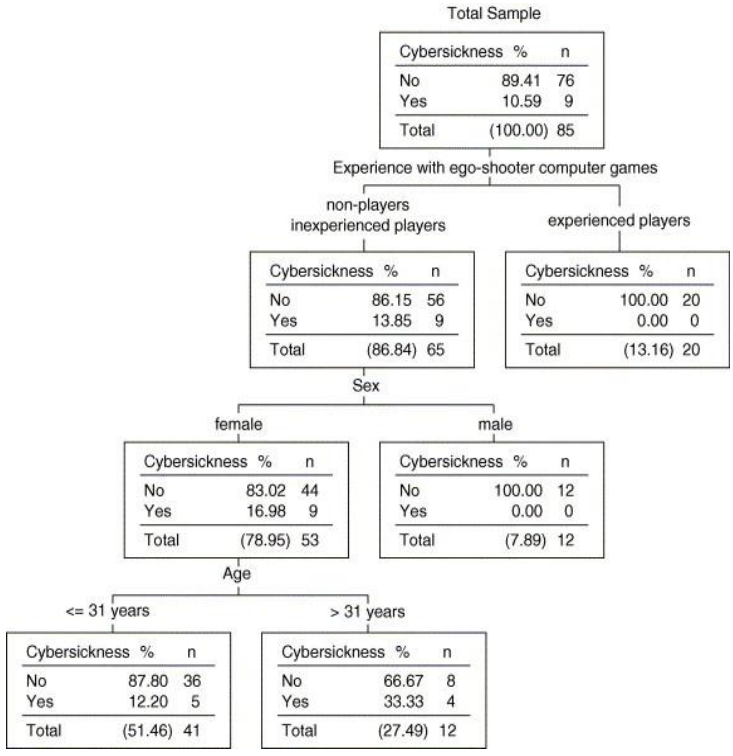


Figure 2 C&RT tree

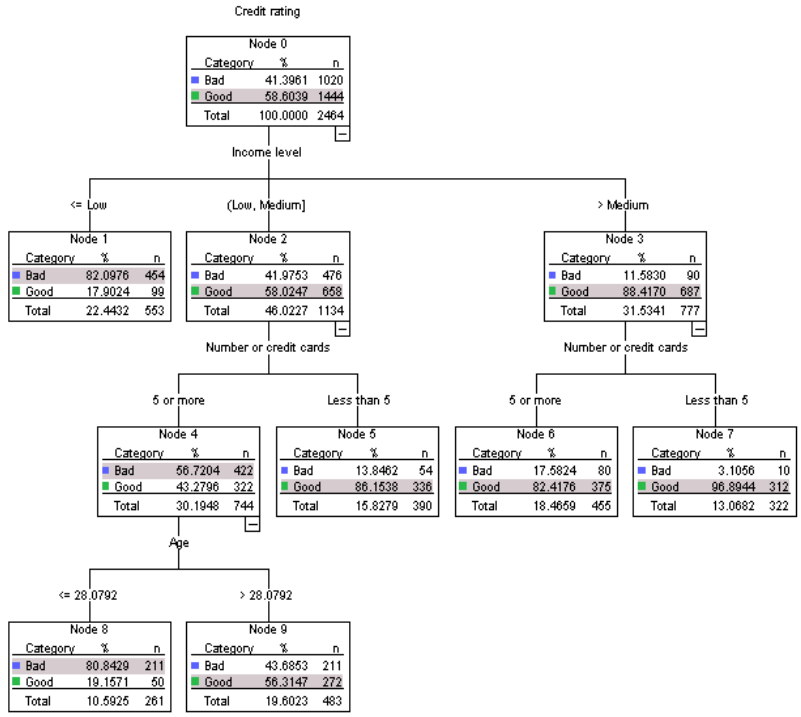


Figure 3 CHAID tree

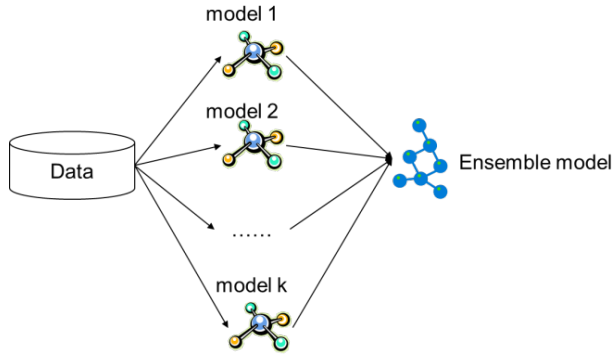


Figure 4 Ensemble Model [6]

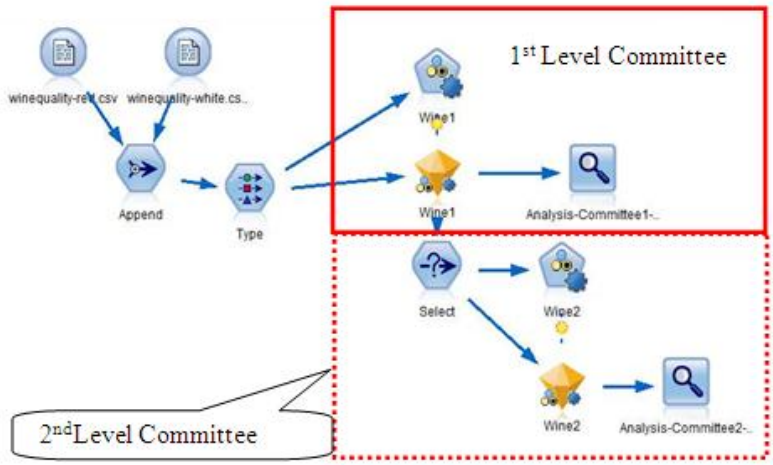


Figure 5 The purposed model of 2-levels ensemble of classifiers.

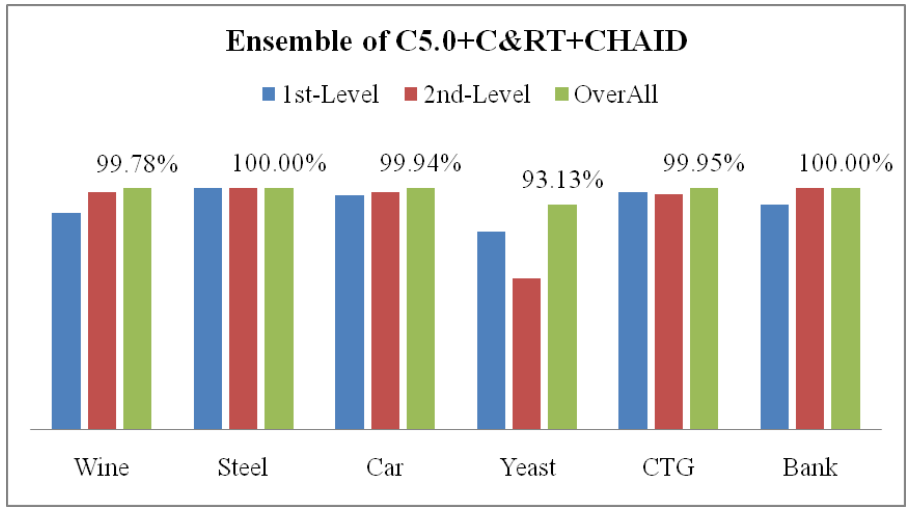


Figure 6 Comparison of performances of 2-levels ensemble tested on 6 data set

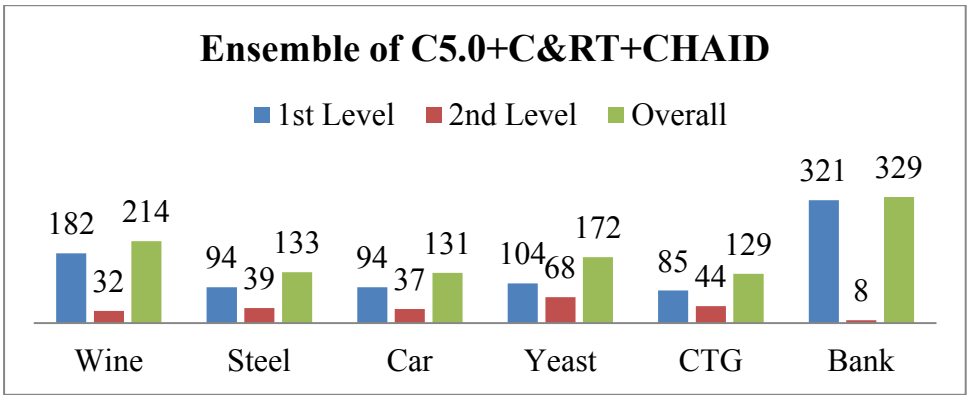


Figure 7 Comparison of time performances of 2-levels ensemble tested on 6 data set in term of execution time (second)