RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS No. 104
Business Informatics 16 2010

Janina A. Jakubczyc
Wrocław University of Economics, Wrocław, Poland
janina.jakubczyc@ue.wroc.pl

THE INTERPRETABILITY OF CONTEXTUAL CLASSIFIER ENSEMBLE

Abstract: While decision trees are usually claimed to be easily interpretable, one cannot say this about decision tree ensembles. Dominant feature of decision tree ensembles is the loss of an interpretability. The aim of this paper is to overcome this issue of comprehensibility by using contextual classifier ensemble.

Keywords: decision tree, tree ensembles, contextual classifier ensembles, interpretability of classifier ensembles.

1. Introduction

The problem concerns interpretability of decision tree ensembles which are limited or discarded compared to a single decision tree. Such a problem exists when we are interested not only in effectiveness of classifier ensemble but also in discovering, learning and understanding relations between phenomena under consideration. Here we are dealing with such a situation. The need for interpretation of tree ensembles is important in the area of economic and social human activity. The reason of this is the lack of consensus which can indicate one of the many existing theories that explains best the socio-economic phenomenon.

The problem of interpretability is important also in the light of the knowledge usefulness. The utility of discovered knowledge is impossible when there is no trust and no acceptance of the user. Moreover, users often construct models to gain insight into the problem domain rather than to achieve an accurate classifier only, as write Van Assche and Blockeel [2007]. Comprehensive knowledge may be exploited by a user to protect from undesired events, or may give the possibility to take an action that fosters the desired behavior.

What does the interpretability stand for and what is causing lack of interpretability of decision tree ensembles? To find an answer for these questions we start from the interpretability concept.

The formal definition of interpretability is possible on the basis of formal disciplines such as mathematical logic. In other areas of research, we use the informal and intuitive definitions. Such a situation occurs inter alia in the area of knowledge dis-

covery from data and in economics and sociology, which are areas of interest. Common definition of interpretability according to D. Ward [2007, p. 11] is as follows: "interpretability reflects the ease with which the user may understand and properly use and analyze the model (data, information)". To particularize this description, we can say that interpretability is determined by the adequacy of used definitions and terminology (this aspect is beyond the scope of our interest yet), user knowledge and his perceptual-cognitive capabilities. Perceptual-cognitive capabilities translate into the complexity and the representation and analysis of the model.

Now we can look for the reasons that have increased complexity of the ensemble model, have made illegible representation and cause great difficulties in model analysis. In our opinion there are three reasons. The first one is the number of base classifiers in decision tree ensemble, that can reach even several dozen, which effectively prevents the analysis of such a complex model.

The second concerns the ways the base classifiers are generated (data set manipulation), which make it impossible to distinguish them [Gatnar 2008; Dietterich 2000]. What is the difference between two decision trees created on the basis of two random samples at a similar level of accuracy of classification? Which one is better and why? Without additional knowledge about each of classifiers answers to these questions are impossible. The third one is the way the ensembles are created, in which the qualification is determined by the level of diversity and classification accuracy of classifier [Dietterich 2000]. This approach fails when the number of potential base classifiers has identical or very similar levels of diversity and efficiency. Then we can say that the selection is random.

Contextual classifier is the proposal in which the basic criterion for generation of classifiers is known or identified are contexts in which classification task can be seen. Contextual classifier represents all contextual situations for each context. Each sample that is the basis for classifier creation is assigned the meaning resulting from the context. This approach determines the maximum number of base classifiers. It gives the possibility to distinguish single classifiers by the contexts in which they describe the investigated phenomenon and it can enrich our knowledge and understanding of the contexts found.

To introduce our idea for this problem solution we first introduce in section 2 selected ideas that support tree ensemble interpretability. The brief characteristic of contextual classifier ensemble, as a proposal that can support interpretability of single and ensemble decision trees, we present in section 3. In section 4 we give an example of creating interpretable decision tree ensemble.

2. The selected support ideas for comprehension of tree ensembles

The support of interpretability to our best knowledge is seen as two-folded issue. The first one is concerned with the ways of looking inside "black box" (ensemble of classifiers). The second applies to simplification of base decision trees and the cardinality of base classifiers in the ensembles.

Analyzing tree ensemble one can gain information including: variable importance, effects of variables on predictions, intrinsic proximities between cases, clustering observation according to proximities measure, scaling coordinates based on the proximities, outlier detection [Breiman 2009].

The similar way of coping with interpretability represent Friedman and Popescu [2008] analyzing influence of rules (they can be generated from decision trees of ensemble) on individual predictions, selected subsets of predictions, or globally over the entire space of joint input variable values. Extensive analysis gives the possibility to assess respective input variables globally, locally in different regions of the input space, or at individual prediction points. They have presented techniques for automatic identifying of those variables that are involved in interactions with other variables, the strength and degree of those interactions, as well as the identities of the other variables with which they interact.

The opinion about improvement of interpretability by looking inside "black box" confirms also Meinshausen [2009]. His proposition of "node harvest" is trying to reconcile the two aims of interpretability and predictive accuracy by combining positive aspects of trees and tree ensembles. This approach consist in the prediction of new observations as the weighted average of the mean responses across all these nodes. It works the better the more rules are included in tree ensemble.

To conclude the ways of looking inside "black box" we can say that enrichment of additional synthetic information about the ensemble tree gives the possibility to significantly better comprehension of the problem under research. The analysis concerns only information about variables (attributes) and relations that is included in generated models. This approach prefers ensembles of great complexity. This opinion confirms for example Breiman [2009], who claims that: "complex classifier can yield a wealth of interpretable scientific information about the prediction mechanism and the data".

The latter approach is to increase interpretability by increasing simplicity of base and ensemble classifiers. Nock [2002] has proposed algorithm WIDC (Weak Induction of Decision Committees) which relies on results about partition boosting, ranking loss boosting, and pruning. It gives the possibility of conducting two sorts of pruning: optimistic and pessimistic. While optimistic pruning outperforms other algorithms in the light of the size of formulas obtained, pessimistic pruning tends to achieve a more reasonable tradeoff, with high accuracies on small formulas.

The opinion that shallower trees with fewer nodes are far easier interpretable than existing ensemble methods using decision trees as weak classifiers, represents Zimmermann [2008].

There exist data sets where the pre-pruning or pruning can be detrimental to the performance of the classifier by too short description, maybe abandonment of interesting information or too high level of model generality.

About the difficulties of quantification of the complexity of decision trees write A. Buja and Y.S. Lee [2001]. In their opinion these difficulties come, inter alia, from

the fact that interpretability does not necessarily determine less complex tree and vice versa. These authors showed that even unbalanced tree with a large depth (e.g., 9 and 13) can be easily interpretable, when discovered dependencies are monotonic.

In the approach presented above we need simple versus complex ensemble. Hence these two approaches cannot be merged. So there is a need in advance to know what approach of supporting interpretability we are going to apply.

In spite of pruning and shallowing tree ensembles, there is an idea of merging all base decision trees into one decision tree. The merged tree combines the best features of base trees from ensemble and expresses the same function as a voted ensemble [Mulvaney, Phatak 2003]. But the results are not better than single decision tree.

Another way presents Szpunar-Huk [2006], who has built rule set from classifier ensemble. Unlike the Mulvaney and Phatak's proposal, her method improves generation abilities of tree ensemble besides the reduction of ensemble complexity. In many cases a single aggregated model can be still quite complex to comprehend.

The presented ideas of interpretability support of tree ensembles can be outline briefly as follows:

- less complex models
  - pruning and shallowing base classifiers,
  - transformation of all the base classifier into one model;
- looking inside "black box", as some synthetic view across all base classifiers.

Two factors that support interpretability, as was showed above, are: simplicity and synthetic knowledge about classifier ensemble. The main disadvantage of this is the fact that we are forced to choose one of them only. It seems that additional knowledge about classifier ensemble may bring more interpretability by indicating the attributes of given importance, frequency of some relations, and so on. It can directly be justified by knowledge and experience of users. It can be treated as suggestion of some discovered relations. But one may claim otherwise.

These solutions in our opinion enrich our knowledge about tree ensembles but this knowledge is insufficient. Both the synthetic measurement and relationship patterns are fragmentary knowledge and its usefulness is not certain. We look for an approach that will generate simple models (simple base tree, few instead of many base classifiers) and gives additional information about each base classifiers.

We refer to infological concept of knowledge [Stefanowicz 2009], which is represented by three elements: information, context and experience. The context is the way one sees the information. It gives directly the indication to interpretability. It seems almost obvious that without the context the information cannot be comprehended. The knowledge without the context and experience is just a systematic set of information. Usually the context of found solutions is given by the user who refers to own experiences and knowledge.

Instead of relying only on the user's context, which is limited to possessed knowledge and gained experiences, there is a possibility to find the contexts in learning sets [Turney 1993]. The found contexts can be used as criteria to build more comprehensive classifier ensembles.

3. Characteristic of the contextual classifier ensemble (CCE)

The ways of generating base decision tree classifiers concern manipulation of learning set. In the area of interest is some kind of random sampling [Dietterich 2000; Dzeroski, Zenko 2004]. The generated classifiers are typically combined by majority or weighed voting scheme. In spite of the demonstrated effectiveness of classifier ensembles, this approach gives the rise to the following problems:

– the number of sets which need to be created in order to have a guarantee of a better description,
– number of base classifiers in ensemble,
– interpretation of differences between various files with learning data (interpretation of single base classifiers),
– explanation of results,
– understanding the impact of the classifier ensemble on the classification results.

On the contrary, contextual approach is using intentional and purposeful way to create base classifiers. It gives the possibility to utilize more information that is included in data sets or outside data set (additional information). The basis for creating base classifiers is the context of given classification problem. The context can be internal, if contained in data set, or external, if it can be found outside of dataset. The context is the criterion for creating learning files for base classifiers. The number of contexts in which some problem can be perceived is finite, so it results in known number of base classifiers. The different context distinguishes base classifiers and gives the interpretability of each single base classifier.

The algorithm for creating contextual classifier ensemble (for more details see [Jakubczyc 2007]):

a) build decision tree on the basis of the entire learning set (basic attributes and irrelevant attributes),

b) context identification: create decision tree for each decision attribute, taking into consideration only irrelevant attributes (context-sensitive attributes from basic attributes; context attributes from irrelevant attributes),

c) context qualification: identify pairs "contextual/context-sensitive" of attributes that can be used to the partition the learning set (according to the assumed level of classification accuracy),

d) build contextual base classifiers for each selected contexts as a compound of decision tree generated for each learning subset of selected context,

e) combine base classifiers into contextual classifier ensemble [Jakubczyc 2007a].

The main advantage of presented algorithm is deliberate and not random way of creating classifiers and using a wider range of information contained in the data. Hence the interpretability can increase significantly.

4. Intelligence of client bank as an example of interpretability

Banks in general need to analyze and estimate their clients' behavior. Thus it would be useful for the bank management to predict if the client is active or non-active. If the client is non-active, there is an indication to take pro-active action to the client to keep him as a stable bank client for the future and do not let him switch the bank. The customer behavior is described by 36 dimensional data (6 are nominal and 30 are real). The associations of the data with particular customer as well as the real meaning of each of presented 36 features are matter of confidentiality and are not revealed. The bank is providing 24 000 representation data for a competition. 12 000 are marked as A (active) and 12 000 are marked as N (non-active) clients [http://neuron.tuke.sk/competition2/upload.php].

As we will see, the idea of contextual classifier ensemble allows an interactive user participation at every stage of creating classifier ensemble.

The identification of contexts has been conducted on the basis of attributes within the decision tree model. Each of these decision attributes has been explained by irrelevant features. As a result, it turned out that all decision attributes are context-dependent. We identified nine contexts that were represented in the form of decision trees. In Table 1 we present indentified contexts, classification accuracy for contextual classifiers and general classifier (built on the basis of entire learning set).

Table 1. The characteristics of the contextual situations and contextual classifiers

Context-dependent attributes	Attributes of contextual situations	Number of components of contextual classifiers	Classification accuracy		
			A	N	Total
B16_E	B15-e, B36-c	13	**79**	**75**	77
B21_N	B28-m, B22-a, B33-y	12	**81**	**72**	**77**
B27_M	B28-m, B14-e,	15	83	71	77
B30_I	B24-y, B29-c	6	**78**	**74**	**76**
B31_N	B32-a, B26-c, B34-y, B14-e, B19-e, B33-y, B3-y	34	78	77	78
B25_C	B19-e, B26-c	10	80	72	76
B11_E	B19-e, B4-y	9	**78**	**73**	75
B12_E	B14-e, B20-ia	8	**78**	**74**	76
B13_E	B14-e, B33-y	13	80	73	77
General			81	67	74

Now we can conduct two-folded analysis of discovered context. The first one is on the basis of some objective criteria: simplicity, similarity, efficiency, and it can look as the paragraph below.

For the description of the context for the most cases, two attributes are sufficient. The exception is situation B31_N that needs for its description seven attributes. In addition, it should be noted that there appear similarities between description of contextual situations that can be identified on the basis of attributes that describe contextual situations. One can, for example, recognize similar pairs of situations: B12_E and B13_E with the common attribute B14_e, B11_E and B25_E with the common attribute B19_e.

Similarities can be the basis for combining contextual situation in some specific way. Number of components of contextual classifiers is of big importance, because it indicates the complexity of contextual classifiers. As we can see, the cardinality of contextual situations is high.

The second fold of analysis is directed to the user. The user can look at identified contexts and learn about them also on the basis of conducted analysis. He can evaluate contexts in the light of his knowledge and experience taking into account their comprehension and novelty, and his expectations. Then the user can choose the ones he is interested in and his choices may be different from the above ones. The user also can postpone a decision until the ensemble is created. Moreover the user can give the appropriate names of analyzed contexts.

Classifier ensemble can be made in three different ways. The contextual classifier may contain all generated classifiers, only contextual classifiers, the selected classifiers, or their fragments (of contextual situations) (for more detail see [Jakubczyc 2007b]).

The selection of candidates for the base classifiers was conducted on the basis of three criteria: classification accuracy, in particular class N, the complexity of the contextual classifier (the number of classifiers included in the contextual classifier), and similarities between the contextual situations which can reduce the number of base classifiers.

Base classifiers should have the highest classification accuracy, the lowest complexity and contextual situations should be different. Since there is no measure of diversity that would enable a clear choice, one should take into account possible imperfection of such selection [Gatnar 2008].

The most effective contextual classifier is B31_C. However, due to extensive fragmentation it was abandoned as the base classifier. The rationale for this decision is that the classifier B16_E is about 68% less complex (13 to 34), and its classification accuracy is lower only by 1-2%. From three pairs of similar contextual situations: B13_E and B12_E, B21_N and B27_M and B25C_C and B11_E, taking into account the complexity of the contextual situations and classification accuracy we chose the following classifiers: B12_E, B21_N and B11_E. The final base classifier becomes B30_I.

At this stage we present process of base classifiers choice to the user and we assume that his decision about list of potential base classifier is the same. So contextual classifier was created from the following classifiers: B11_E, B12_E, B16_E, B21_N, B30_I. We have also created all possible ensembles of three classifiers from

the selected ones. The best four of them were included in Table 2. For the sake of comparison possibility we have also created two complex classifiers: the first containing all the classifiers and the second containing only the contextual classifiers. The classification accuracy of contextual classifier ensembles is shown in Table 2.

Table 2. Classification accuracy of contextual ensembles

No.	Classes / Ensembles	N	A	Total	Number of base classifiers
1	B11_E, B12_E, B16_E, B21_N, B30_I	77.53	82.47	80.00	5
2	All	77.77	86.74	82.28	11
3	All contextual	77.41	84.66	81.03	10
4	B12_E,B16_E,B21_N	77.02	82.52	79.77	3
5	B13_E,B16_E,B30_I	76.98	82.98	79.98	3
6	B12_E,B13_E,B16_E	76.45	83,05	79.75	3
7	B12_E,B21_N,B30_I	76.39	81.80	79.09	3

On the ground of the analysis of contextual classifier ensembles contained in Table 2, we can say that the accuracy of classification has the values at the similar level. In this case, one can choose classifier ensemble with fewer base classifiers, not losing much in efficiency, and may count on additional profit from interpretability.

Each of the classifiers 4-7 is almost equally effective and only slightly inferior to the more complex classifiers 1-3 (up to 1.38 for class N). It is therefore possible not only to reduce the complexity of the classifier ensemble, but also it is possible to choose classifier ensemble that is more comprehensive and expected for the user.

So the task of classification can be described by a number of contextual classifier ensembles. The selection of the proper ensemble, at this stage, can be left to the user or the expert again. This fact is not of no importance since a user often expects to find relationships between certain attributes only [Breiman et al. 1984].

It should be noted that in the selection of a classifier the key role play contextual situations for which the classifiers were created. The user is able to evaluate the identified contextual situation in terms of comprehension and extension of knowledge.

5. Summary

The proposal to apply contextual classifier ensemble seems very promising. In the light of human limitation of possessed knowledge and experience it gives additional knowledge about data in the form of contexts and contextual situations. Besides the limited number of contexts, in which given phenomenon can be perceived, the achieved classification accuracy is acceptable. This approach supports interpretability of classifier tree ensemble also by user interactivity during the process of building contextual classifier ensembles, for example: user can qualify identified situation

according to his own expectation, domain knowledge and comprehension, can create ensembles using own criteria, can expand knowledge about the problem under research.

The introduced proposal still needs inquisitive research. The main directions of this research concern: identification of context, looking for some ways to shorten context descriptions, and determination the areas of application.

References

Breiman L. (2009), *Looking inside the black box*, http://stat-www.berkeley.edu/users/breiman/wald2002-2.pdf (8.12.2009).

Breiman L., Freidman J.H., Olshen R.A., Stone C.J. (1984), *Classification and Regression Trees*, Wadsworth International Group, Belmont, CA, pp. 203-215.

Buja A., Lee Y.S. (2001), Data mining criteria for tree-based regression and classification, [in:] *Proceedings of the 7th International Conference on Knowledge Discovery in Databases*, ACM, New York, pp. 27-36.

Dietterich T.G. (2000), Ensemble methods in machine learning, [in:] *First International Workshop on Multiple Classifier Systems*, Eds. J. Kittler, F. Roli, *Lecture Notes in Computer Science*, Springer Verlag, New York, pp. 1-15.

Dzeroski S., Zenko B. (2004), Is combining classifiers with stacking better than selecting the best one?, *Machine Learning*, Vol. 54, pp. 255-273.

Friedman J.H., Popescu B.E. (2008), Predictive learning via rule ensembles, *The Annals of Applied Statistics*, Vol. 2, No. 3, pp. 916-954.

Gatnar E. (2008), *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa.

Jakubczyc J.A. (2007a), Contextual classifier ensembles, [in:] *Business Information Systems*, Ed. W. Abramowicz, *Lecture Notes in Computer Science*, Vol. 4439, Springer, Berlin–Heidelberg.

Jakubczyc J.A. (2007b), Predykcja migrujących klientów firmy telekomunikacyjnej z wykorzystaniem złożonego klasyfikatora kontekstowego, [in:] *Systemy wspomagania organizacji*, Eds. T. Porębska-Miąc, H. Sroka, Wydawnictwo Akademii Ekonomicznej, Katowice.

Meinshausen N. (2009), *Node harvest: Simple and interpretable regression and classification*, working paper http://arxiv.org/abs/0910.2145 (19.12.2009).

Mulvaney R., Phatak D.S. (2003), A method to merge ensembles of bagged or boosted forced-split decision trees, *IEEE Transaction on PAMI*.

Nock R. (2002), Inducing interpretable voting classifiers without trading accuracy for simplicity: Theoretical results, approximation algorithms, and experiments, *Journal of Artificial Intelligence Research*, Vol. 17, pp. 137-170.

Stefanowicz B. (2009), Informacja i wiedza, [in:] *Aspekty informatyzacji organizacji*, Ed. A. Nowicki, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 55, Informatyka Ekonomiczna 13, Wydawnictwo Uniwersytetu Ekonomicznego, Wrocław, pp. 381-391.

Szpunar-Huk E. (2006), Classifier building by reduction of an ensemble of decision trees to a set of rules, [in:] *Proceedings of the Computational Intelligence for Modeling, Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce (CIMCA-IAWTIC'06)*, IEEE Computer Society, Washington, DC, pp. 144-155.

Turney P.D. (1993), Robust classification with context-sensitive features, [in:] *Proceedings of the Sixth International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, Edinburgh, Scotland, June, pp. 268-276.

Van Assche A., Blockeel H. (2007), Seeing the forest through the trees: Learning a comprehensible model from an ensemble, *Lecture Notes in Computer Science*, Vol. 4701, Springer, Berlin–Heidelberg, pp. 418-429.

Ward D. (2007), *Data and Metadata Reporting and Presentation Handbook*, http://www.oecd.org/dataoecd/46/17/37671574.pdf (11.03.2010).

Zimmermann A. (2008), Ensemble-trees: Leveraging ensemble power inside decision trees, *Lecture Notes in Computer Science*, Vol. 5255, Springer, Berlin–Heidelberg, pp. 76-87.

MOŻLIWOŚCI INTERPRETACYJNE KONTEKSTOWEGO KLASYFIKATORA ZŁOŻONEGO

Streszczenie: modele drzew decyzyjnych uważa się za łatwe do interpretacji, ale nie można tego powiedzieć o zespołach drzew decyzyjnych. Utrata możliwości interpretacyjnych stanowi istotne ograniczenie w ich zastosowaniach. Celem tego artykułu jest przedstawienie propozycji tworzenia zespołów drzew decyzyjnych według kryterium kontekstu, dającej możliwość zwiększenia interpretacyjności zespołów klasyfikacyjnych.