Title	A cognitively inspired rule-plus-exemplar framework for interpretable pattern classification
Author(s)	Sit, Wing Yee; Mao, K. Z.
Citation	Sit, W. Y., & Mao, K. Z. (2012). A cognitively inspired rule-plus-exemplar framework for interpretable pattern classification. 2012 15th International Conference on Information Fusion (FUSION), pp. 1188-1195.
Date	2012
URL	http://hdl.handle.net/10220/19143
Rights	

A Cognitively Inspired Rule-Plus-Exemplar Framework for Interpretable Pattern Classification

Wing Yee Sit, K. Z. Mao
School of Electrical and Electronic Engineering
Nanyang Technological University
Singapore
sitw0002@e.ntu.edu.sg, ekzmao@ntu.edu.sg

***Abstract*—While the generalizability of classifiers receive much attention in research, interpretability is often neglected. This paper proposes a rule-plus-exemplar classification framework based on ideas in cognitive psychology. The classification process is interpretable and intuitive, and also generalizes well. It can perform better than other interpretable methods such as decision trees, for both interpolative and extrapolative generalization.**

Keywords-pattern classification; ensemble classifiers; cognitive-based classifier; extrapolation; generalization; interpretability

I. INTRODUCTION

Generalizability and interpretability are both important aspects of pattern classifiers. The generalizability of a classifier, learned from a limited number of training data, refers to its ability to correctly classify unseen testing data which are different from the training dataset. This is usually measured by the classification accuracy, and is often the main focus and interest of research.

Interpretability, on the other hand, is the ability of the classification process to be understood by humans, especially the users of an application utilizing a pattern classifier. The classifier should therefore be represented in an intuitive and comprehensible way. This limits the complexity of the classifier, leading to a tradeoff between interpretability and accuracy [1]. As a result, the interpretability aspect of classification is often neglected.

There is no denying the importance of classifier accuracy, especially in applications where the classifier is part of an automated process. For example, in the identification of spam mail, it is important that the legitimate mail is received while the spam is filtered out. The interpretability of the classification process is not important in such applications.

However, in some applications, even though accuracy is still desired, interpretability plays an equally important role. For example, determining whether a patient has a certain disease cannot be done entirely by a classifier. The classification should only serve as a decision support for the medical expert. Determining whether a subject is behaving suspiciously or whether an identified object is a real target may also not be an entirely automated process. Wrong identification could lead to unnecessary panic or a serious lapse in security. Human judgment and expert knowledge may still be required, so a simple yes-no answer from a classifier serves no purpose in aiding the decision. Instead, the process of classification should be interpretable and comprehensible, such that the user can draw from this information and make an appropriate decision.

In this paper, we will focus on both generalizability and interpretability of classifiers. Even though the former has been given much attention in research, only the interpolative aspect is considered in most cases. To determine the generalizability of a classifier, the accuracy is found on a testing dataset which is different from the training data. This dataset is usually generated through randomly ordered splits of a main dataset, or cross validation, bootstrapping, etc. The train and test sets are thus partitions of a same dataset, and are close to each other. The classification accuracy then only reflects the classifier's interpolative generalizability. Yet generalization should also encompass the correct response to samples that lie further from the training data, i.e. outside the feature space covered by the training data.

In order to improve the generalizability, especially extrapolative generalizability, without forgoing the interpretability aspect of classifiers, we make use of ideas from cognitive psychology. The resulting framework makes use of global and local classification methods, which are combined based on the principles of a rule-plus-exemplar cognitive categorization model.

The problem is further elaborated in Section II. Section III provides a look into categorization models in cognitive psychology, and then explains the classification framework, detailing how the ideas from cognitive psychology are relevant and lead to the combination presented. Experiments and results are presented in Section IV, with discussions and analysis before concluding in Section V.

II. PROBLEM AND MOTIVATION

A. Interpretability of pattern classifiers

While there exist many pattern classification methods, most of them are not intuitive or easily comprehensible by humans. However, interpretability is a highly desired property of classification in many applications.

Although many black-box models such as SVM or ensemble classifiers outperform more comprehensible methods

like decision trees in terms of classification accuracy, they face a lack of user receptiveness in certain applications [2]. This could be due to the nature of the application area, such as safety critical domains like airline or air traffic control, power stations, medical and health informatics and security related domains [3, 4], where classification should support the human decision making process. In areas such as medical diagnosis where the classification will influence patient treatment [5], there are regulations that restrict the use of black-box classifiers.

The importance of interpretability goes beyond improving user receptiveness in various applications. Reference [3] described a number of benefits, which includes the following:

- Providing a clear view of the classification process for user evaluation on classifier suitability
- Verification of whether the classifier satisfies requirements of the larger system on which it is integrated
- Acting as a data exploration tool by explaining the complex mappings between data and their output
- Construction of a knowledge base

To provide an interpretable method to pattern classification, we will focus on classifiers that are comprehensible, in particular decision trees. Rules can be extracted from these trees, of the form "*If…Then…*" which are intuitive and natural to users.

There are some generic rule extraction methods that can be used on black-box models [3, 6]. However, they are not widely accessible, and many rule sets extracted only provide an insight to the workings of the classifier and not match its classification. Furthermore, for comprehensible rules that are extracted, they can be presented in the same "*If…Then…*" form. Thus they can be applied in the same framework proposed in this paper, in place of the decision trees.

B. Generalizability of pattern classifiers

Compared to interpretability, generalizability receives far more attention in research. It is the main evaluation means for comparing the performance between classifiers. It is presented as the classification accuracy of the classifier on testing data, which is different from the training data used for learning. It is important that this measure is calculated on separate testing data since a high accuracy performance on the training data could be a result of overfitting rather than a reflection of generalizability.

Although the accuracy measure reflects how well the classifier is likely to perform on unseen data, the way it is usually calculated does not necessarily reflect the classifier's generalizability. This is because the testing data are not located too far from the training data. The two sets are usually generated as partitions of a randomly ordered set of the original dataset, or separate cross-validation datasets. As a result, they are likely to be located within the same region in feature space, even if they represent different samples. This means that the generalization performed here is mainly interpolative. It only shows how well the classifier can classify samples that lie among the training data.

Yet generalizability encompasses more than just interpolation. Humans are able to generalize well based on very limited instances, to correctly identify and categorize unseen instances that are significantly different from before. They are also able to generalize regularities from past experiences to apply correctly in new experiences [7]. This is a reflection of the extrapolative capability of humans. Such ability is crucial to learning and knowledge generation.

Compared to interpolative capability, the extrapolative element of generalization is less comprehensibly investigated. Some methods define a decision region [8, 9], which covers the samples that can be adequately classified, and allow the classifier to return "unclassifiable" otherwise. This could lead to an excessive number of samples labeled as "unclassifiable" by a conservative classifier. Some measures take this into consideration to strike a balance between overgeneralization and overfitting. This is done by weighting such errors with an appropriate cost, along with other errors such as false positives and false negatives [10, 11].

Labeling a test sample as "unclassifiable" when it is far from the training data is a natural and intuitive response. This is because of the assumption that the training data should be sufficiently representative of the actual concept, including the testing data. While this is a fair assumption to make, it is not always satisfied in real applications.

Firstly, training data may not always be easily available or obtainable. They may be limited in numbers by nature, such as disease cases in medical diagnosis or natural disasters occurrences. They could also be costly to obtain, such as clinical tests performed or presence of underground resources. With a limited training data, the whole concept to be learnt cannot be fully represented. Yet the testing data could easily fall in other regions of the concept not covered by the training data.

Secondly, the assumption may be violated due to a drift in the testing data as a result of changes in the environments. The classifier may be trained on a large set of training data, but as a part of an operating system, testing data may be provided sequentially over time. Any changes in the environment, such as surrounding temperature that affects the temperature reading or lighting that affects color features, could cause perturbations to the testing data. Wear and tear of machines or sensors could also cause changes to the features of the testing data, resulting in an overall shift of the region beyond that covered by the training data. This may not be as drastic as in concept drifting, but still translate to increasing distance from the training data.

The problem arises when the testing sample is located far from the training data. This is because the classifier is no longer reliable, as shown in Figure 1. The samples from each class are denoted by 'o' and '+' and two different decision tree classifiers were trained on the samples. The region mapped to each class is colored in the figure with different shades. The boundary between classes can thus be seen, with (a) depicting the boundary obtained using CART and (b) using C4.5. Using the different decision tree classifiers, the accuracies on the

testing data are the same. But outside the region covered by the training data, the boundary between classes extends in very different ways. Besides, the boundary between the classes extends in a counter intuitive manner.

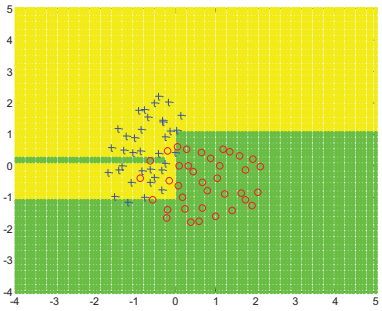


Figure 1. Extension of boundary formed by decision trees beyond covered region (a) CART (b) C4.5

To truly improve generalization, the classifier should be able to handle such data to a certain extent in the above situations, until resources are available for retraining or recalibration of machines.

With both aspects given equal importance, the goal of this paper is to provide a classification method that can retain interpretability while improving both the interpolative and extrapolative generalizability.

III. COGNITIVELY INSPIRED CLASSIFICATION FRAMEWORK

A. *Ideas from cognitive psychology*

Cognitive psychology is the study of how humans think and how information is stored and processed. Categorization is one of the major processes, covering the building of concepts and applying them on new instances. In order to appropriately and reliably classify testing data that are further from the training data, we can look into ideas from cognitive psychology. Although pattern classification methods can deal with much higher data complexity, humans are far more superior in generalizing and adapting to changing concepts. An understanding of how humans perform categorization can provide a useful framework for improving the extrapolative generalizability of classifiers. In addition, categorization models in cognition are naturally intuitive and comprehensible to us, which can help uphold interpretability.

Two main categorization models have been described – rules and exemplars. Early work found that humans performed categorization using rules, which are created from the encountered instances, and there are also experiments that illustrate their use [12-14]. Later experiments show that they were insufficient to explain many phenomena that were observed, and exemplars were shown to be stored and used for comparison during categorization [15-17]. Here, exemplars are the instances or samples that have been encountered. Due to the complementary nature of the two methods, hybrid systems were proposed [16-18], in which categorization was carried out using a representation consisting of both rule and exemplar subsystems.

B. *Application of cognitive ideas in classification*

Exemplar models came into the picture as rules were unable to explain a variety of experimental observations as well as to cope with ill-defined categories [19]. Otherwise, rules tend to be the preferred means for categorization as it represented abstracted knowledge, functioned faster and also did not require recollection of specific instances.

Many classifiers, like the rule subsystem in cognitive psychology, represent an abstraction of knowledge. This knowledge is extracted from all the training data through an attempt to optimize certain measures on this data. These measures may be mean squared error, entropy or separation, among others. Such classifiers are global methods which construct the model for the whole problem space defined by the training data, but does not consider beyond. Thus when a testing sample lies outside this space, it is classified based on information across the whole region but which may not be relevant to the sample itself.

Local methods, on the other hand, work like the exemplar subsystem. They focus on sub regions of the whole space. This is most apparent in the k nearest neighbor (kNN) method. The exemplar subsystem, while differing between various models and implementations, is similarity-based and the influence of the most similar exemplars is the strongest. Like kNN, even though it may not perform optimally in the region covered by the training data, it can provide an intuitive classification outside the region. This is because of the focus placed on the most relevant information in the training data, which are the training samples closest to the testing sample. Besides, they can cope with exceptions better than global methods.

Therefore, the complementary nature of the rule and exemplar subsystems in cognitive psychology is reflected in pattern classification as well. Within the region covered by the training data, global methods are preferred, not necessarily because of their performance, but because of their efficiency and abstraction. They provide a higher level knowledge of the concept, and do not require tedious comparisons. Outside this region, they can be complemented by local methods. This brings us to the classification framework for improving generalizability.

C. *Global and local methods simulating rule and exemplar subsystems*

As explained in the earlier section, the complementary nature of rule and exemplar subsystems in cognitive

categorization can be reflected by the global and local classifiers. For better interpretability, decision trees are used for the global methods. This is also in line with the categorization model, as rules can be easily extracted from decision trees that provide equivalent classification. The local classifier is implemented using kNN, which is in line with the exemplar categorization model. The key then lies in putting the two subsystems together in a cognitively intuitive way to improve generalization.

D. Improving extrapolative generalizability

Let us explain the combination in terms of global and local classifiers. It has been shown in the earlier section that in order to improve extrapolative generalizability, a local classifier can be used to classify test samples that are far from the training data, i.e. lying outside the region covered by the training data. This helps to focus the classification on information most relevant to the test sample outside. This problem translates to the question of when the testing sample is too far. In other words, we need to identify the region that can be covered by the training data. This has been covered in [20].

The region is easily identifiable visually in two or three dimensional space. But when the problem extends to higher dimensions, this region has to be defined based on the training data using some evaluation function. This evaluation function should have higher values for samples close to the training data, and diminish in value as samples move further away. It cannot be calculated as distance to the center point of the training data due to different possible distributions of the training data, as well as possible multi modal data. This function can be represented as

$$T(x) = \frac{1}{M}\sum_{i=1}^{M} F_i(x), \tag{1}$$

where M is the number of training samples. Each $F_i(x)$ should be a function that has largest value at the point x, but decrease away from it. It can be implemented as a Gaussian function centered at training sample s_i, and is defined as

$$F_i(x) = exp\frac{-\|x - s_i\|^2}{2{\sigma_i}^2}. \tag{2}$$

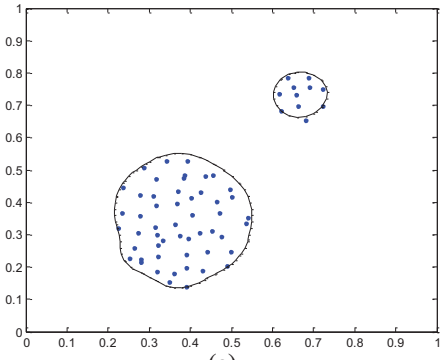
Each σ_i determines the spread of the Gaussian function at that training sample s_i. Values too large will result in over-regularization while values too small will result in under-regularization of the function to the covered region. It is thus computed separately for each training sample by pegging to the distance to neighbors.

$$\sigma_i = \|s_i - N_3(s_i)\| \tag{3}$$

where

$$N_3(s_i) = 3^{\text{rd}} \text{ nearest neighbor of } s_i. \tag{4}$$

This ensures that the spread is large enough such that any sample located between training samples can be covered by the Gaussian functions of at least one of its nearest training samples.


(a)

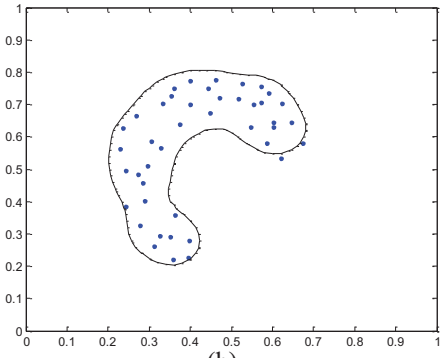

(b)

Figure 2. Region identified to be covered by training samples

The evaluation function T(x) provides a measure of how close a sample is to the training data. Defining the region covered by the training data will require a threshold for the T(x) values, such that the region can be defined as

$$R = \{x: T(x) > \tau\} \tag{5}$$

for some threshold τ. This threshold can be obtained directly from the T(x) values for the training data. However, due to the possibility of outliers, it is not set to the minimum T(x) value, but instead to the fifth percentile. This value performs adequately for most cases, and the region identified can be seen in Figure 2 for different datasets. It can also be adjusted according to the dataset.

The algorithm for determining if a test sample lies in the region covered by the training sample is shown in Figure 3.

M = number of training samples
For *each training sample* s_i
  $N_3(s_i)$ = *3rd nearest neighbor of* s_i
  $\sigma_i = \|s_i - N_3(s_i)\|$
  *i*th *element of* T_{tr}, $T_{tr}(i) = T(s_i)$
End
Find threshold $\tau = percentile(T_{tr}, 5)$
Given test sample x
If $T(x) \geq \tau$,
  x *is in covered region R*
End

Figure 3. Algorithm for finding region threshold parameter

Although the evaluation function itself is not directly interpretable, it has an intuitively comprehensible significance to it. The function merely translates to a single "*If…Then…*" rule to determine whether a sample is inside the region covered by the training data.

E. Improving interpolative generalizability

The exemplar subsystem can support rules not only for generalization outside the covered region, but also inside. In particular, the exemplar subsystem supports the rules when the categories are ill-defined. This is when the rules are inadequate for classification, resulting in many exceptions to the rules.

Exception handling is one of the main limitations to rules. They are better represented using exemplars, and thus the exemplar subsystem can come into play here. The decision tree essentially partitions the feature space into different parts, each corresponding to a rule and assigned a class label. Based on the training data, the performance of each rule can be determined, and those which have poor accuracy can be identified. These correspond to the regions where the categories are ill-defined and require support from exemplars.

However, due to the handling of ill-defined regions by decision trees, finer partitions of these regions will be created. Each partition could cover very few training samples, but due to overfitting, results in very high training accuracy among these regions. These regions can be merged together if they both have little coverage and are subpartitions of a larger region. There may also be other regions with low coverage, but if they were split from a larger region with high coverage, they would not represent an ill-defined region.

The subpartitions with low coverage are hence merged. A total coverage of not more than 5 samples has been found to perform reasonably well. This is because due to higher dimensions of datasets, fine partitions generally cover very few samples. These resulting regions, along with other poor rule regions, will constitute the ill-defined regions. Test samples falling in these regions will switch from decision tree to kNN classification.

```
For each decision tree node,
  If children of current node are leaves AND
      their total coverage ≤ 5
  Merge nodes
  End
End
Translate leaves to rules
For each rule,
  If error of rule ≥ ρ,
    Label rule as poor performing
  End
End
```

Figure 4. Algorithm for finding poor performing rules

The problem now lies in identifying the rules which are inadequate in performance. This requires a cut-off point in the rule error, above which a rule is deemed to be performing poorly. The threshold ρ can be adjusted based on the datasets, although a value of 20% generally suits most data. This also tallies with the subpartition merging criteria, as any merged region would have a minimum error of 20%. The regions found can all be clearly described using "*If…Then…*" rules, thus retaining interpretability. This procedure is described in Figure 4.

F. Classification framework

Using the rule-plus-exemplar framework from cognitive psychology, the resulting framework is proposed to improve both interpolative and extrapolative aspects of generalization. The overall framework is summarized in Figure 5.

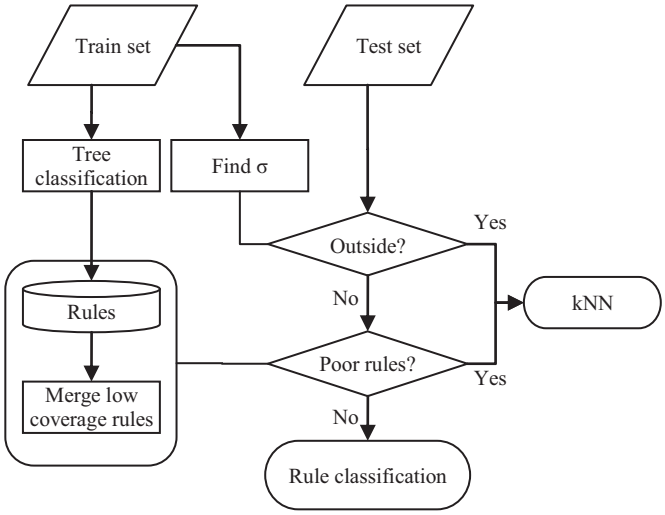


Figure 5. Flow process of rule-plus-exemplar classification framework

Based on the training data, the region covered by the training data is identified using the parameter σ. The decision tree classifier is also trained, and rules are extracted. Those which are inadequate are identified through the parameter ρ. During testing, a sample is checked against the covered region. If it lies outside, it will be classified using kNN. Otherwise, it will be checked against the region covered by inadequate rules. If it lies inside such a region, kNN is again used for classification. Otherwise, it does not belong to any of the special cases requiring support from exemplars, so the rules are adequate for classification.

G. Other existing methods

Our proposed classification framework can be understood as a type of ensemble. Many ensemble classifiers which use different base classifiers combine the prediction result using various voting schemes [21], Naïve Bayes combination, etc. These are fusion ensemble methods which apply all the classifiers on the test sample before combining them.

Reference [22] also suggested combining global decision trees with local instance based KNN, but various base classifiers are generated for both methods using different parameters, and the classification results are combined through voting, stacking or grading.

The proposed framework falls under the class of selection ensemble methods [23], in which each classifier specializes in a different region of the feature space. Some methods identify the regions of competence [24] explicitly through regular splitting or clustering, or implicitly using estimates such as local

accuracy estimate [25]. However, such methods are not interpretable, and are also based on improving only the interpolative generalizability.

Fusion of classifier outputs also does not facilitate interpretation of the classification process. This is especially true for another line of ensembles, which is the mixture of experts. The use of a gating network is a black-box model which is not comprehensible.

IV. Experiments and Analysis

The cognitively inspired classification framework aims to improve interpolative and extrapolative generalization while retaining interpretability. Extrapolative generalization requires testing data that are not only different from training data, but also lying outside the region covered by the training samples. Such testing data will be constructed from commonly used datasets.

A. Datasets for training and testing

Datasets for experiments in this section are obtained from the UCI Machine Learning Repository. The data is normalized to zero mean and unit variance, and the samples can be classified into one of two classes. The OQ dataset is obtained by using only the 'O' and 'Q' classes from the Letter Recognition Dataset.

To ensure that there is testing data lying outside the region covered by the training data, we separate the dataset into internal and external datasets. This process is described in Figure 6. The T(x) values are computed for all the samples, and then ranked within each class. The top 20% of samples with smallest T(x) values are found, and τ_c is found for each class c. The threshold τ is given by the smallest τ_c value, such that the external set is made up of any sample with T(x) value smaller than τ. This ensures that the external set is not too large and will not cover all the samples of one class.

For *iteration = 1,…,100*
  Use 70% of the full dataset
  Calculate $T(x)$ *for all the samples*
  For *each class c,*
    Rank the samples by $T(x)$ *values:* $t_1, \dots, t_{m_c}$
    $\tau_c = t_{\lfloor 0.2 m_c \rfloor}$
  End
  $\tau = min_{c=1,2}(\tau_c)$
  $External\ set = \{x : T(x) \leq \tau\}$
  $Internal\ set = \{x : T(x) > \tau\}$
End

Figure 6. Algorithm for finding external and internal datasets

The internal set is further separated into a testing set with 20% of the samples, and a training set containing the rest of the samples.

B. Extrapolative generalizability

Firstly, we demonstrate the feasibility of the rule-plus-exemplar model for improving extrapolative generalizability outside the region covered by the training data. The results in Table I are based on the internal and external testing sets.

TABLE I. Error rates inside and outside covered region

Datasets	Inside covered region		Outside covered region			
	CART	*C4.5*	*CART*	*C4.5*	*Exemplars for extrapolation*	
					CART	*C4.5*
Australian	16.53%	19.16%	24.54%	30.14%	19.65%	19.71%
Bands	34.77%	36.46%	42.54%	36.15%	38.98%	38.98%
Blood Transfusion	30.04%	29.07%	33.14%	29.22%	25.74%	25.70%
Credit Approval	15.86%	19.53%	21.81%	24.93%	17.12%	17.14%
Glass-2	29.58%	30.00%	35.64%	34.81%	34.30%	34.30%
OQ	6.03%	9.17%	19.35%	23.09%	9.72%	9.74%
Parkinsons	19.31%	19.65%	14.71%	11.23%	8.97%	8.44%
Pima Indians	28.80%	29.64%	37.11%	35.59%	36.03%	36.03%
Sonar	28.96%	33.38%	46.72%	52.17%	32.12%	32.12%
Spect	24.88%	26.95%	4.51%	4.59%	1.61%	1.61%
Spectf	28.45%	32.58%	13.71%	48.90%	0.95%	0.95%
Statlog Heart	24.18%	28.93%	32.50%	33.77%	20.66%	20.70%
Vertebral	20.21%	22.72%	15.07%	15.87%	10.17%	10.14%
Wdbc	7.73%	8.05%	9.33%	9.77%	5.43%	5.45%
Wpbc	32.08%	37.43%	39.73%	53.70%	22.46%	22.66%

Classification is performed without any prior knowledge that testing data are located outside the covered region. They are determined based on the process described in Figure 3.

The box outlined in bold in Table I compares the performances of two most popular decision tree methods CART and C4.5, comparing their performances inside and outside the region covered by the training data. In most datasets, the error rates on testing data outside the covered region are significantly higher than on testing data inside. The deterioration is especially serious in the Sonar and Wpbc datasets, where the classification outside is worse than chance. By correctly identifying such samples and directing to exemplar-based classification, the error rates were reduced.

For some datasets, the extension of the boundary under decision tree classification may be aligned with natural class separation, thus the exemplar use may not be better. However, it is generally improved for most datasets.

As the covered region is identified independently of the decision tree classifier, the last two columns of Table I give largely the same result. This is because most of the samples have been correctly identified to be outside the covered region, and hence is directed to kNN, using k = 3.

C. Interpolative generalizability

The cognitively inspired framework supports high level rules with exemplars not only for samples outside the region covered by training data. For regions where categorization is ill-defined, instance based classification is also used. This is meant to improve the interpolative generalizability of

TABLE II. INTERPOLATIVE GENERALIZATION IMPROVEMENT

Datasets	CART error rates		C4.5 error rates	
	Rules only	*Exemplar use for poor rules*	*Rules only*	*Exemplar use for poor rules*
Australian	16.53%	15.51%	19.16%	17.09%
Bands	34.77%	32.31%	36.46%	35.50%
Blood Transfusion	30.04%	29.10%	29.07%	28.28%
Credit Approval	15.86%	14.68%	19.53%	17.02%
Glass-2	29.58%	27.48%	30.00%	29.49%
OQ	6.03%	5.25%	9.17%	7.47%
Parkinsons	19.31%	16.71%	19.65%	17.91%
Pima Indians	28.80%	27.33%	29.64%	28.79%
Sonar	28.96%	26.81%	33.38%	32.75%
Spect	24.88%	23.24%	26.95%	24.83%
Spectf	28.45%	27.10%	32.58%	31.13%
Statlog Heart	24.18%	22.02%	28.93%	26.31%
Vertebral	20.21%	20.57%	22.72%	22.31%
Wdbc	7.73%	7.05%	8.05%	7.29%
Wpbc	32.08%	30.76%	37.43%	36.43%

interpretable rules. This is valid for samples that are within the covered region.

Table II shows the error rates on the internal testing set using CART and C4.5. These are compared with the error rates when kNN is used for test samples lying in regions corresponding to poor performing rules. The procedure is that which is described in Section III-E.

The use of exemplars for ill-defined regions improves the classification in almost all cases. The robustness and results of the rule-plus-exemplar classification framework shows that the ideas from cognition are valid in pattern classification as well. The use of instance based classification can improve the generalizability of rules over a wide range of problems, while maintaining a good level of interpretability.

D. Overall generalization improvement

The results of the previous experiments are indicative of the improved interpolative and extrapolative generalization capabilities. This can now be verified in an overall system, implemented as described in Figure 5.

Indeed, the use of exemplars improves the overall generalizability of the classification. The test set in this case consists of not only samples within the region covered by the training data, but also samples outside. The classification remains comprehensible to users, as it is based on the rule-plus-exemplar framework that is cognitively intuitive to us.

E. Discussion and analysis

The main goal of the proposed classification framework is to retain interpretability while improving generalization. However, due to the tradeoff between comprehensibility and accuracy, there is a limit to the level of accuracy that can be attained. Thus, it is only fair to compare the performances of the classification framework with other methods of comparable comprehensibility, such as decision trees.

Nevertheless, we can show how the system fares in comparison with black-box classification models. Support vector machines are commonly used for pattern classification and generally produce good accuracy.

Classification is performed on the overall testing set containing samples inside as well as outside the region covered by training samples. Errors rates are expected to be lower in SVM than the rule-plus-exemplar model, as it allows higher complexity in the classification. However, SVMs are not easily interpretable to users [26].

Nevertheless, the error rates achieved are sometimes comparable. The rule-plus-exemplar result from using CART was better than C4.5 for the following datasets. The rule-plus-exemplar error rates were 19.51% compared to 18.07% in SVM for Australian dataset, 17.39% compared to 16.37% in SVM for Credit Approval dataset, and 14.21% compared to 13.88% in SVM for the Parkinsons dataset. These performances can be achieved despite the rule-plus-exemplar framework being far more interpretable than the black-box SVM model.

Another question that may come to mind is the use of exemplars. If they are better than rules at handling ill-defined categories and samples far from the training set, and are also comprehensible, it is natural to ask why exemplars are not simply used without rules. The reason is that rules represent a higher level knowledge that has been extracted from training

TABLE III. OVERALL GENERALIZATION

Datasets	CART error rates		C4.5 error rates	
	Rules only	*Rules + exemplars*	*Rules only*	*Rules + exemplars*
Australian	20.50%	19.51%	24.61%	23.79%
Bands	37.73%	35.68%	36.40%	35.45%
Blood Transfusion	31.86%	30.13%	29.07%	28.47%
Credit Approval	19.06%	17.39%	22.43%	20.90%
Glass-2	31.78%	29.71%	31.80%	31.43%
OQ	11.92%	10.57%	15.32%	14.48%
Parkinsons	17.02%	14.21%	15.68%	15.38%
Pima Indians	32.50%	31.20%	32.29%	32.17%
Sonar	36.38%	33.44%	41.23%	38.29%
Spect	14.47%	13.12%	15.53%	14.75%
Spectf	21.40%	19.40%	40.40%	38.96%
Statlog Heart	27.98%	25.10%	31.23%	28.73%
Vertebral	17.72%	16.63%	19.43%	18.70%
Wdbc	8.38%	7.36%	8.74%	8.24%
Wpbc	35.98%	32.30%	45.68%	41.63%

data. They also allow generalization of concepts, which is important in gaining insight into problems. Furthermore, rules are faster for performing classification, as they do not require tedious comparisons with individual training samples, and also do not require storage of all the training data. The use of rules for most of the cases is thus preferred, even though exemplars can be utilized when rules are inadequate.

V. CONCLUSION

The classification framework in this paper demonstrates the feasibility of classifying using a framework similar to that in cognitively psychology. It enables better generalization while retaining the interpretability. This is demonstrated through the experimental results. Such a framework lays the foundation for further work. In addition to better interpretability, the framework also facilitates the incorporation of expert knowledge. It can be implemented as rules, or under the exemplar representation for higher priority in classification. Such knowledge can be added and removed as needed, without retraining of the classification system.

In improving interpolative generalizability, the current method for identifying poorly performing rules is based directly on the training data. However, as decision trees are prone to overfitting, the rules can be better evaluated using a separate validation data set when the training data is sufficiently large.

In this paper, the region of feature space covered by the training data is identified in a generic method that is applicable regardless of the decision tree method used for modeling the rule subsystem. There are other rule induction methods which can also be used. Some may not assume a default class, but instead directly identify the regions that are covered by the training data. This rule-plus-exemplar framework remains valid and applicable.

REFERENCES

[1] T. Van de Merckt and C. Decaestecker, "About Breaking the Trade Off Between Accuracy and Comprehensibility in Concept Learning " presented at the Proceedings of the Workshop on Comprehensibility in Machine Learning, IJCAI-95, Montreal 1995.

[2] J. Huysmans and B. Baesens, "Using Rule Extraction to Improve the Comprehensibility of Predictive Models," Katholieke Universiteit Leuven Leuven, Belgium 2006.

[3] R. Andrews, *et al.*, "Survey and critique of techniques for extracting rules from trained artificial neural networks," *Knowledge-Based Systems,* vol. 8, pp. 373-389, 1995.

[4] Z. Xing, *et al.*, "Extracting interpretable features for early classification on time series " presented at the Proceedings of the 11th SIAM International Conference on Data Mining (SDM'11) Phoenix, Arizona, USA 2011.

[5] G. Fung, *et al.*, "Rule Extraction from Linear Support Vector Machines " presented at the Proc. ACM SIGKDD '05, 2005.

[6] N. Barakat and A. P. Bradley, "Rule extraction from support vector machines: A review," *Neurocomputing,* vol. 74, pp. 178-190, 2010.

[7] Y. Munakata and R. C. O'Reilly, "Developmental and computational neuroscience approaches to cognition: the case of generalization," *Cogn. Stud,* pp. 76-92, 2003.

[8] H. N. A. Pham and E. Triantaphyllou, "An application of a new meta-heuristic for optimizing the classification accuracy when analyzing some medical datasets," *Expert Syst. Appl.,* vol. 36, pp. 9240-9249, 2009.

[9] O. Melnik, "Decision Region Connectivity Analysis: A Method for Analyzing High-Dimensional Classifiers," *Mach. Learn.,* vol. 48, pp. 321-351, 2002.

[10] H. N. A. Pham and E. Triantaphyllou, "A meta-heuristic approach for improving the accuracy in some classification algorithms," *Comput. Oper. Res.,* vol. 38, pp. 174-189, 2011.

[11] H. Pham and E. Triantaphyllou, "The Impact of Overfitting and Overgeneralization on the Classification Accuracy in Data Mining," in *Soft Computing for Knowledge Discovery and Data Mining*, O. Maimon and L. Rokach, Eds., ed: Springer US, 2008, pp. 391-431.

[12] L. J. Rips, "Similarity, typicality, and categorization," in *Similarity and analogical reasoning*, ed: Cambridge University Press, 1989, pp. 21-59.

[13] S. W. Allen and L. R. Brooks, "Specializing the Operation of an Explicit Rule," *Journal of Experimental Psychology: General,* vol. 120, pp. 3-19, 1991.

[14] M. A. Erickson and J. K. Kruschke, "Rules and exemplars in category learning," *Journal of Experimental Psychology: General,* pp. 107--140, 1998.

[15] R. M. Nosofsky, *et al.*, "Rules and exemplars in categorization, identification, and recognition," *Journal of Experimental Psychology: Learning, Memory, and Cognition,* pp. 282--304, 1989.

[16] J. K. Kruschke and M. A. Erickson, "Learning of rules that have high-frequency exceptions: New empirical data and a hybrid connectionist model " presented at the Sixteenth Annual Conference of the Cognitive Science Society Hillsdale, N.J.: Erlbaum, 1994.

[17] A. Vandierendonck, "A parallel rule activation and rule synthesis model for generalization in category learning," *Psychonomic Bulletin & Review,* vol. 2, pp. 442-459, 1995.

[18] R. M. Nosofsky, *et al.*, "Rule-Plus-Exception Model of Classification Learning," *Psychological Review,* vol. 101, pp. 53-79, 1994.

[19] S. C. McKinley and R. M. Nosofsky, "Investigations of exemplar and decision bound models in large, ill-defined category structures " *Journal of Experimental Psychology: Human Perception and Performance,* vol. 21, pp. 128-148, 1995.

[20] W. Y. Sit and K. Z. Mao, "Cognitively Inspired Classification for Adapting to Data Distribution Changes," presented at the IEEE Evolving and Adaptive Intelligent Systems 2012 Madrid, to be published.

[21] R. Battiti and A. M. Colla, "Democracy in neural nets: voting schemes for classification," *Neural Netw.,* vol. 7, pp. 691-707, 1994.

[22] D. Baumgartner and G. Serpen, "Comparative performance evaluation of global-local hybrid ensemble," *International Journal of Hybrid Intelligent Systems,* vol. 8, pp. 59-70, 2011.

[23] B. V. Dasarathy and B. V. Sheela, "A composite classifier system design: Concepts and methodology," *Proceedings of the IEEE,* vol. 67, pp. 708-713, 1979.

[24] L. I. Kuncheva, "Combining pattern classifiers : methods and algorithms," ed New York: J. Wiley, 2004, pp. 189-202.

[25] K. Woods, *et al.*, "Combination of multiple classifiers using local accuracy estimates," in *Computer Vision and Pattern Recognition, 1996. Proceedings CVPR '96, 1996 IEEE Computer Society Conference on*, 1996, pp. 391-396.

[26] V. Cherkassky and F. Mulier, "Noninductive Inference and Alternative Learning Formulations," in *Learning from Data: Concepts, Theory, and Methods* ed: John Wiley & Sons, Inc., 2007, pp. 467-498.