УДК 519.68

О ПРОГРАММНОЙ РЕАЛИЗАЦИИ И АПРОБАЦИИ АЛГОРИТМА DFBSA СИНТЕЗА ЭМПИРИЧЕСКОГО РЕШАЮЩЕГО ЛЕСА

Ю.Ю. Дюличева

Таврический национальный университет им. В.И. Вернадского,
факультет математики и информамтики
пр-т Вернадского, 4, г.Симферополь, Крым, Украина 95007
e-mail: *dyulicheva@hotbox.ru*

Abstract

Pioneering recognition technology *Forest Based Learning (FBL)* software implementation based on new estimation methods, *D*ecision *F*orest *B*uilding *S*equencing *A*lgorithm and pruning strategy is proposed.

Введение

Многие прикладные задачи, возникающие при создании информационных систем, связаны с распознаванием свойств, типов, классов или уровней качества объектов. В частности, таков является задача распознавания болезнетворных вибрионов, возбудителей кишечно-желудочных болезней, оценивая перспективности инвестирования, диагностики технических объектов. Сложность получения начальных данных - числовых описаний объектов - приводит к тому, что во многих случаях имеется их описание только в терминах первичных свойств - признаков - в виде наборов значений логических переменных.

Среди методов распознавания, активно используемых в мировой практике, наиболее распространены методы, основанные на решающих деревьях и нейросетевых технологиях. Нейросетевые алгоритмы менее приспособлены для решения задач с качественными переменными и не позволяют получать логические описания классов объектов в явном виде.

Анализ последних исследований и публикаций свидетельствует о появлении широком применении ряда программных комплексов, основным инструментом, для принятия решений в которых, служат решающие деревья (РД). Наиболее известными среди них являются CART [1], C4.5. [2]. В то же время известен эффект «перенастройк» РД на обучающие данные (overfitting) [3], что привело к появленю различных методов коррекции структуры дерева [4], [5] (редукция [6, 7, 8, 9], наращивание отдельных вершин (grafting) [10] и т.д.).

Предложенный автором данной статьи алгоритм синтеза эмпирического решающего леса *DFBSA* [12, 13, 14] удачно сочетает как стратегию редукции, так и теоретическую возможность возврата назад с целью выбора «информативных» признаков во внутренние вершины решающего дерева.

Целью настоящей работы является обоснование с практической точки зрения, полученных в [11]-[16] теоретических результатов. Предлагаемая программная реализация информационно-распознающей системы *Forest Based Learning (FBL)* предназначена для решения задач обучения и распознавания объектов на основе технологии решающих деревьев и принципиально нового приема оценивания, редукции отдельных классификаторов и организации последовательных процедур построения решающей среды.

Актуальность разработанного программного комплекса подтверждается востребованностью высокоточных алгоритмов обучения распознаванию для широкого класса приложений, и в частности, для решения задач, связанных с обнаружением и распознаванием болезнетворных вибрионов; *его новизна* заключается в принципиально новом подходе к построению распознающей процедуры, неизвестной в мировой литературе и обеспечивающей снижение ошибки распознавания.

1. Алгоритм DFBSA обучения и процедуры распознования

Алгоритм $DFBSA$ *(Decision Forest Building Sequencing Algorithm)* синтеза эмпирического решающего леса (ЭРЛ) удачно сочетает стратегию редукции, отсекая ветви, ранг которых превышает заданное пороговое значение; возможность возврата (отката) для построения нового разбиения по системе признаков, вообще говоря, не участвовавших ранее в обучении; процесс синтеза очередного дерева, прежде всего, на тех объектах, для которых предыдущее дерево не способно сформировать «простое» правило, описывающее их класс; процесс «достройки» РД на оставшихся объектах таблицы обучения. Алгоритм $DFBSA$ основан на построении области отказа для каждого РД, входящего в лес. Область отказа представляет собой интервал, описывающей конъюктивную закономерность, соответствующую ветви РД некоторого ранга и содержащий объекты обучающего множества, которые не попадают в лист некоторой ветви допустимого ранга, а попадают в лист некоторой ветви допустимого ранга, а попадают в терминальную вершину специального вида (cut), означающую отказ от обучения на таких объектах.

Последовательный синтез деревьев леса направлен на формирование области отказа очередного дерева и сужение совокупной области отказа - пересечений областей отказа деревьев листа. Если в результате синтеза совокупная область отказа окажется пустой, то построен корректный решающий лес, поскольку для каждого объекта из обучающего множества найдется хотя бы одна конъюнкция допустимого ранга, входящая в описание класса, которому этот объект принадлежит Рис.1 демонстрирует схему каждого РД эмпирического решающего леса.

В качестве некоторых критериев остановки синтеза ЭРЛ можно предположить следующие [12, 13, 16]:

- в результате синтеза очередного РД леса совокупная область отказа не сужается, т.е нельзя построить допустимое разбиение на интервалы допустимого ранга, содержащие объекты из совместной области отказа (нарушается условие корректности леса);

- ограничение на число деревьев леса.

Можно предположить различные процедуры распознавания на основе эмпирического решающего леса. Вкратце опишем некоторые из них.

Для распознавания предоставляется описание объекта в виде булевого вектора длины n.

Процедура распознавания с переходами по «ссылкам». Указатель устанавливается на первое дерево. Согласно значениям признаков осуществляется «прохождение» соответствующей ветви и определение терминальной вершины. Если метка листа - класс, процесс распознавания объекта завершен; иначе (лист имеет специальную метку cut) указатель указатель устанавливается на следующее дерево. Результат этой

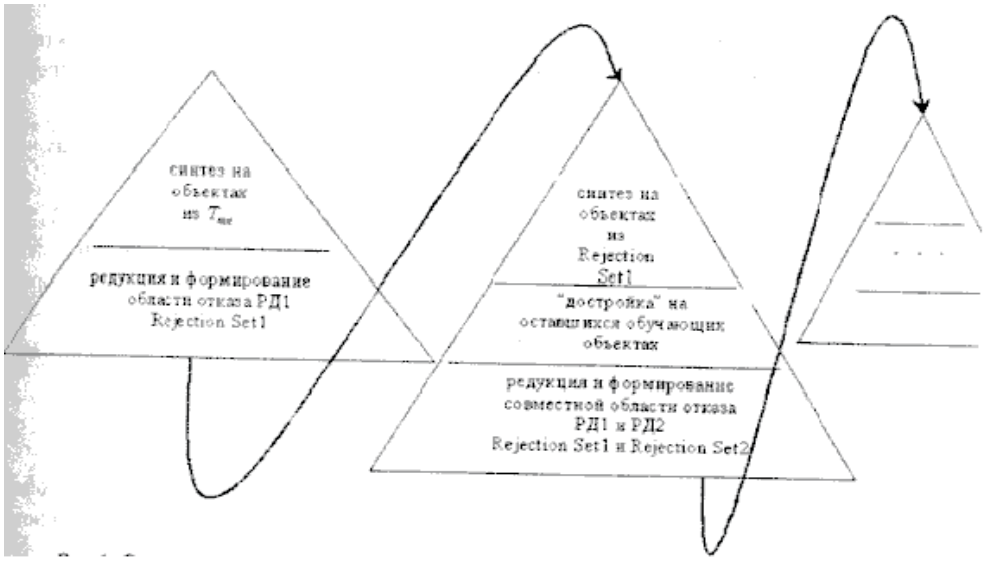

Рис. 1. Схема синтеза каждого решающего дерева эмпирического леса

процедуры определяется порядком обхода деревьев леса. Переход на новое дерево соответствует классификации объекта, вообще говоря, по новой системе признаков. Предыдущий отказ от решения и ссылка на следующее дерево не оказывают никакого влияния на дальнейший процесс распознавания.

Процедура распознавания на основе наиболее «компонентной» ветви эмпирического леса. Указатель устанавливается на первое дерево. Согласно значениям признаков осуществляется «Прохождение» ветви и определение соответствующего листа.

Если лист содержит метку класса, осуществляется переход по указателю на следующее дерево в поисках ветви наименьшего ранга, оканчивающейся листом с меткой класса, которому принадлежат объект. Таким образом, объект распознан ветвью, имеющей наилучшую статистическую оценку.

Процедура распознавания на основе «голосования» ветвей ЭРЛ. Каждое дерево эмпирического решающего леса можно рассматривать как эксперта, принимающего решение в своей области «компетентности» (совокупности непересекающихся интервалов допустимого ранга). Каждый эксперт-дерево при поступлении описания объекта может либо проголосовать за определенный класс либо воздержаться от голосования (ветвь заканчивается меткой cut). Сначала указатель устанавливается на первое дерево эмпирического леса. Согласно значением признаков осуществляется прохождение ветви и определение соответствующего листа. Если лист помечен меткой класса, эксперт-дерево голосует за этот класс. В результате алгоритма осуществляется просмотр всех деревьев эмпирического леса и подсчитывается число голосов, отданное экспертами в пользу каждого класса. Класс, набравший наибольшее число голосов, определяет метку класс объекта, поступившего для распознавания.

В частности, в программной комплексе FBL, описываемом ниже, реализована процедура распознавания с переходами по «ссылкам».

Все эти процедуры целесообразно применять, когда построен корректный эмпирический лес. В случае некорректного эмпирического леса возможно применение алгебраического корректора. Этот подход был подробно изучен в [15].

2. Реализация и апробация новой технологии Forest Based Learning (FBL)

В данном разделе представлены результаты практического применения алгоритма синтеза решающего леса для классификации базы данных болезнетворных вибрионов, описываемой таблицей обучения, содержащей 365 булевых 153-мерных векторных, разделенных на 5 классов:

Vibrio cholerae 01 серогруппы, биовар eltor серовар Огава;
Vibrio cholerae non 01;
Aeromonas;
Vibrio alginolyticus;
Vibrio parahaemolyticus.

Система FBL позволяет по обучающей информации построить как одно корректное решающее дерево, так и эмпирический решающий лес, ранг которого может быть задан пользователем. Существенным обстоятельством является то, что и дерево и лес используют в процессе синтеза один и тот же критерий ветвления. В системе FBL в качестве критерия ветвления был выбран и реализован D-критерий максимальной отделимости пар разных классов [17]. Алгоритмы синтеза РД и ЭРЛ реализованы

с использованием инструментального средства Delphi в среде Windows 95/98. Система FBL представляет информацию для сравнения характеристик и качества корректным РД и корректным РД и корректным ЭРЛ. В частности, можно получить информацию о числе контрольных объектов, распознанных решающим деревом и эмпирическим решающим лесом, числе объектов из совокупной области отказа на каждом этапе синтеза очередного дерева леса, числе листьев в каждом дереве леса, ранге ветвей РД, на котором были допущены ошибки при распознавании объектов контрольной выборки и описания классов объектов в виде ДНФ.

Рис.2-3 иллюстрируют корректной эмпирический решающий лес, состоящий из 3 деревьев, ранг каждой ветви которых не превышает 5. В таблицах 1-2 для сравнения приведены некоторые сведения о корректном РД (максимальный ранг ветви 7) и корректном ЭРЛ, полученные в одном из экспериментов (310 объектов на обучение, 55 объектов на контроль).

Рис.4 демонстрирует процесс обучения корректного РД и средний процент ошибки на контроле в 150 экспериментах. Как видно из рис.4 средний процент ошибок на контроле, совершаемых корректным РД, не ниже 6%.

В таблице 3 проводится сравнительный анализ качества распознавания корректным РД и корректным ЭРЛ при длинах обучающей и контрольной выборок соответственно 310 и 55.

Таблица 1. Сравнение характеристик корректного РД и корректного ЭРЛ

	число ошибок на контроле	число использованных признаков	номера использованных признаков	число листьев
Корректное РД	5 из 55	19	3,6,15,18,25,29,37,40,43,51, 53,59	21
Корректный ЭРЛ	4 из 55	34		28
РД1		15	6,15,18,25,29,43,51,53,79,99, 114,132,147,150,152	14
РД2		19	3,4,7,11,17,28,52,54,56,59,71, 82,84,107,108,115,130,135,149	13
РД3		-	-	1

Таблица 2. Отличительные характеристики корректного ЭРЛ

Корректный ЭРЛ	Число объектов в области отказа	Номера объектов из области отказа
РД1	14	111,240,243,246,247,248,261,272, 286,290,298,301,310
РД2	5	240,243,246,248,272
РД3	0	-

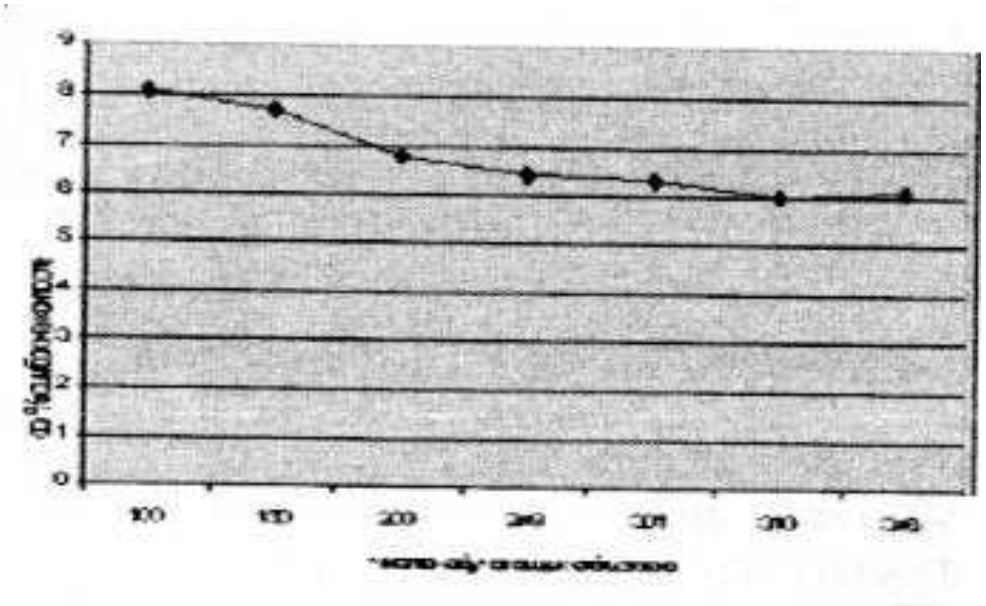

Таблица 3. Качество распознавания объектов на контроле корректны РД и корректным ЭРЛ

	Ср. % ошибки на контроле
РД	6,1
РЛ ранга 5	6,18
РЛ ранга 6	5,5

Приведем описания классов, полученные по корректному РД, для данных таблицы 1.

Описания классов по корректному РД:

Vibrio cholerae 01 серогруппы, биовар eltor серовар Огава ($w1$)

$x_2 5 x_9 9 x_7 9$

Vibro cholerae non ($\omega 4$)

$\bar{x}_3 x_{150} \bar{x}_{152} x_{51} \bar{x}_{114} \bar{x}_{79} \vee x_{43} x_6 \bar{x}_{132} \bar{x}_{99} x_{79} \vee \bar{x}_{144} x_{132} \bar{x}_{99} \bar{x}_{79} \vee x_{18} x_{114} x_{132} \bar{x}_{99} x_{79} \vee x_{29} \bar{x}_{25} x_{99} x_{79}$

Aeromonas ($\omega 6$)

$x_{114}\bar{x}_{79} \vee \bar{x}_{15}\bar{x}_6\bar{x}_{132}\bar{x}_{99}x_{79} \vee \bar{x}_{43}x_6\bar{x}_{132}\bar{x}_{99}x_{79} \vee \bar{x}_{18}x_{114}x_{132}\bar{x}_{99}x_{79} \vee \bar{x}_{29}\bar{x}_{25}x_{99}x_{79}$

Vibrio alginolyticus ($\omega 10$)

$x_{40}\bar{x}_{147}\bar{x}_{53}\bar{x}_{51}\bar{x}_{114}\bar{x}_{79} \vee \bar{x}_{37}x_{147}\bar{x}_{53}\bar{x}_{51}\bar{x}_{114}\bar{x}_{79} \vee x_{53}\bar{x}_{51}\bar{x}_{114}\bar{x}_{79} \vee \bar{x}_{59}x_3x_{150}\bar{x}_{152}x_{51}\bar{x}_{114}\bar{x}_{79} \vee$
$\vee\, x_{15}\bar{x}_6\bar{x}_{132}\bar{x}_{99}x_{79}$

Vibrio parahaemolyticus ($\omega 14$)

$x_{40}\bar{x}_{147}\bar{x}_{53}\bar{x}_{51}\bar{x}_{114}\bar{x}_{79} \vee x_{37}x_{147}\bar{x}_{53}\bar{x}_{51}\bar{x}_{114}\bar{x}_{79} \vee x_{53}\bar{x}_{51}\bar{x}_{114}\bar{x}_{79} \vee \bar{x}_{59}x_3x_{150}\bar{x}_{152}x_{51}\bar{x}_{114}\bar{x}_{79} \vee$
$\vee\, x_{15}\bar{x}_6\bar{x}_{132}\bar{x}_{99}x_{79}$

Для описания классов в виде ДНФ, привычных и полезных при анализе единичных РД, в ЭРЛ фигурирует существенно более сложная конструкция. Прежде, чем приступить к ее описанию, заметим, что процесс принятия решения и его описания - существенно разные вещи. Корректный ЭРЛ определяет разбиение B^n на области, соответствующие классам и, соответственно, однозначные решающие правила (алгоритмические функции классов). Действительно,из определения процедуры распознавания с переходами по «ссылкам», очевидно, что для любого $\bar{x} \in B^n$ либо однозначно определяется его метка $\omega(\bar{x})$ (в этом случае завершится выполнение процедуры распознавания с переходами по «ссылкам») либо ссылка на следующее дерево, причем для корректного дерева в последнем по порядку РД ссылок нет, следовательно, $\bar{x}$получает ровно одну метку.

Пусть $W_i : B^n \to \{\omega_1, \omega_2, \ldots, \omega_l, \triangle\}$- алгоритмическое отображение, определяемое i-м РД леса $\{DT_1, DT_2, \ldots, DT_q\}$,где l-число классов, $\triangle$ - отказ от решения.

Под областью компетентности i-го решающего дерева DT_i будем понимать множество

$$CompTree(DT_i) = \bigcup_{j=1}^{l} \{\tilde{x} \in B^n | \omega_j(\tilde{x}) \neq \triangle\}$$

Очевидно, что имеют место соотношения:

1^0 $\bigcup_{i=1}^{q} CompTree(DT_i) = B^n$

2^0 $\forall i, sCompTree(DT_i) \bigcap CompTree(DT_s) = \emptyset$

Пусть $CompSet(\tau) = U_{i=1}^{\tau} CompTree(DT_i)$, тогда, очевидно $CompSet(1) \subseteq CompSet(2) \subseteq \ldots \subseteq CompSet(q) = B^n$.

Каждое последующее по порядку, определенному алгоритмом принятия решения, РД леса вычисляет отображение W_i только на сужении $D_{i-1} - B^n/CompSet(i-1)$, которое может описано в виде логической формулы $F_{i-1}(x_1, x_2, \ldots, x_n)$ так2ой, что $F_{i-1}(x_1, x_2, \ldots, x_n) = \begin{cases} 1, & \tilde{x} \in D_{i-1} \\ 0, & \tilde{x} \in D_{i-1}. \end{cases}$. Эта формула может быть представлена некоторой ДНФ $D_{i-1} = K_{i-1}^1 \vee K_{i-1}^2 \vee \ldots, K_{i-1}^v$. Если в DT_i, определенно решающее правило $L_1^i \vee L_2^i \vee \ldots \vee L_u^i$ для некоторого класса в виде ДНФ, то для логического описания ω_j следует использовать $(K_1^{q-1} \vee K_2^{q-1} \vee \ldots \vee K_{\beta_{q-1}}^{q-1}) \cdot (L_{1,\omega_j}^q \vee \ldots \vee L_{u_{\omega_j},\omega j}^q)$. Но не смотря

на это, сложные формулы определяют не используемые для вывода решения конъюнкции, а лишь условия применения этих конъюнкций, определяемых областями компетентности.

Построение ДНФ класса по ЭРЛ, как описания решения.

1^0 Взять все ветви РД1, помеченные метками классов, и «расписать» конъюнкции по классам, получая ДНФ $D_1(\omega_j), j = \overline{1,l}$. Записать ДНФ R_1, соответствующую ветвям, помеченным меткой cut (R_1 - описание области отказа). $D_1^0(\omega_j) = D_1(\omega_j)$.

i^0 Пусть построены $D_{i-1}^0(\omega_j), R_{i-1}$. По DT_i получим: $D_i^0(\omega_j) = D_{i-1}^0(\omega_j) \vee (R_{i-1} \wedge D_i(\omega_j))$.

Для построенного ЭРЛ (Рис.2-3), получим описание класса Vibrio cholerae 01 серогруппы, биовар eltor серовар Огава($\omega 1$):

$D_1(\omega_1) = x_{79}x_{99}x_{25}$,

$R_1 = \bar{x}_{79}\bar{x}_{114}\bar{x}_{51}\bar{x}_{53} \vee \bar{x}_{79}\bar{x}_{114}x_{51}\bar{x}_{152}x_{150}$,

$D_1^0(\omega_1) = x_{79}x_{99}x_{25}$,

$x_{11}x_{84}\bar{x}_3x_{56}\bar{x}135 \vee x_{71}x_{82}x_{130}\bar{x}_{56}\bar{x}_{135}$,

$D_1(\omega_1) \vee (R_1 \wedge D_2(\omega_1)) = x_{79}x_{99}x_{25} \vee \bar{x}_{79}\bar{x}_{114}\bar{x}_{51}\bar{x}_{53}x_{11}x_{84}\bar{x}_3x_{56}\bar{x}_{153} \vee$

$\vee \bar{x}_{79}\bar{x}_{114}x_{51}\bar{x}_{152}x_{150}x_{71}x_{82}x_{130}\bar{x}_{56}\bar{x}_{135}$

Заметим, что решение, определяемое ЭРЛ согласно процедуре распознавания с переходами по «ссылкам», всякий раз принимается одной конъюнкцией ограниченного ранга, но, возможно, в условиях отказа предыдущих по порядку РД леса.

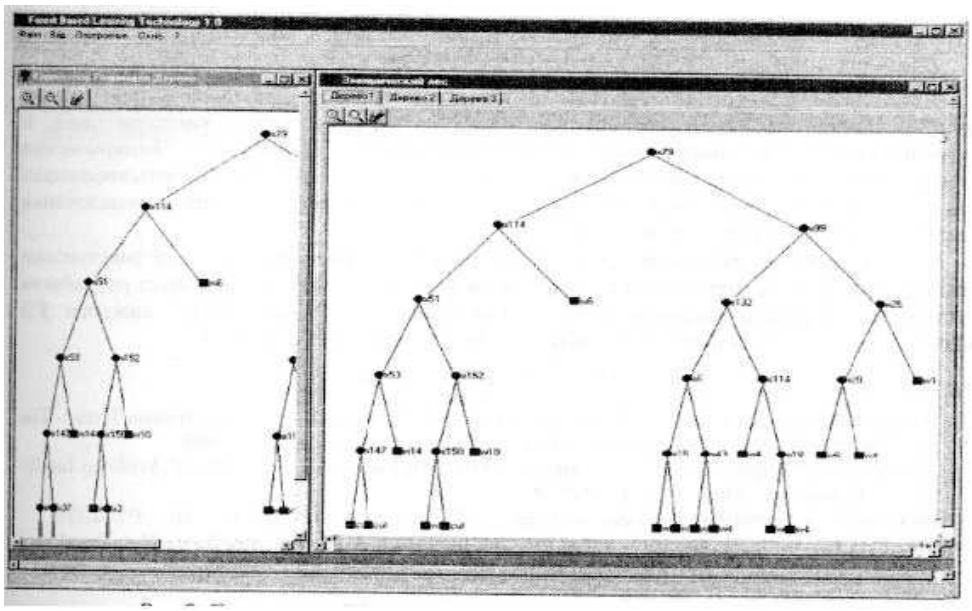

Рис. 2. Построение РД1 и редукция ветвей, ранг которых больше 5

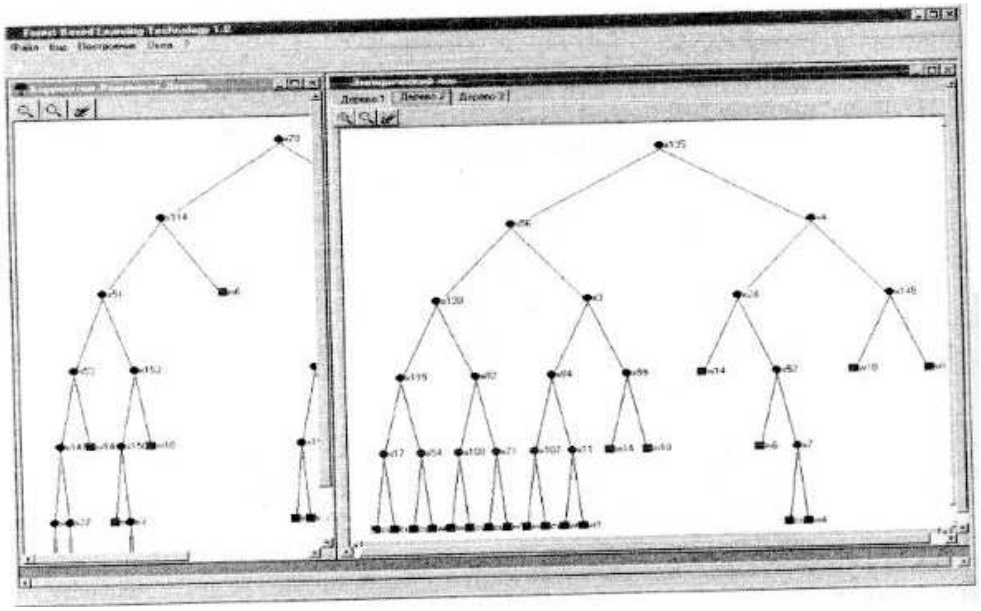

Рис. 3. Построение РД2 на объектах из области отказа РД1 и его достройка на оставшихся обучающих объектах

Заключение

Полученные ранее в работах [11]-[16] *теоретические результаты подтверждены при решении практически важной задачи* эффективного распознавания и формирования описаний классов болезнетворных вибрионов. Эмпирический решающий лес повысил эффективность распознавания объектов, не участвовавших ранее в обучении, по сравнению с одним решающим деревом, при использовании одного и того же критерия ветвления.

Представляется перспективным исследование вопроса оптимальной расстановки ссылок при переходе от дерева к дереву в эмпирическом решающем лесе; разработка стратегий целенаправленного поиска признаков для «достройки» каждого РД эмпирического решающего леса; дальнейшее изучение свойств ЭРЛ.

Список литературы

1. *Breiman L.,Friedman J.H., Olshen R.A., Stone C.J.* Classification and Regression Trees. - The Wadsworth Statistics (Probability Series). Belmont, CA: Wadsworth, 1984.
2. *Quinian J.R.* Improved Use of Continuous Attributes in C4.5 // Journal of Artificial Intelligence Research.—1996.Vol.4 — P.77-90.
3. *Schaffer C. Overfitting Avoidance as Bias* //Machine Learning. - 1993. - 10. - P.153.178.
4. *Kothari R., Dong M.* Decision Trees for Classification: A Review and Some New Results // Pattern Recognition: From Classical to Modern Approaches, S.R. Pal(Eds), World Scientific. – 2001. – Chapter 6. – P.169-184.

5. *Freud Y., Mason L.* The Alternating Decision Tree Learning Algoritm // Proceedings of 16th International Conference on Machine Learning . – Morgan Kaufmann, San Francisco, CA. – 1999. – P.124-133
6. *Esposito F., Malerba D., Semeraro G.* A Comparative Analysis of Methods for Pruning Decision Trees // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 1997. – 19(5). – P.476-491.
7. *Fournier D., Cremilleux B.* A Quality Index for Decision Tree Pruning // Knowledge-Based System. – 2002. – Vol.15. – P.37-43.
8. *Hall L.O., Collins R., Boyew K.W.*, Banfield Error-Based Pruning of Decision Trees Grown on Very Large Data Sets Can Work! // International Conference on Tools for Artificial Intelligence. – 2002. – P.233-238.
9. *Дюличева Ю.Ю* Стратегии редукции решающих деревьев (обзор) // Таврический весник информатики и информатики. – 2002. – №1 – С.10-17.
10. *Webb G.I.* Decision Tree Grafting From the All-Test-But-One Patrition // Proceedings of 16th International Joiint Conference on Artificial Intelligence. – Morgan Kauffman. – San Francisco, CA. – 1999. – P.702-707.
11. *Донськой В.Й., Дюлічева Ю.Ю.* Бінарні розв'язувачі дерева в задачах інтелектуального аналізу інформації // Наукові вісті Націо нального технічного університету України «Київський політехічний Інститут». – 2001. – вип.5. – С.12-18.
12. *Донской В.И., Дюличева Ю.Ю* Индуктивная модель r-корректного эмпирического леса // Труды Междунар. конф. по индуктивному моделированию. – Львов. – 2002. – С.54-58.
13. *Дюличева Ю.Ю* Оценка VCD r-корректного ємпирического леса // Таврический вестник информатики и математики. – 2003 – № 1. – С.30-42.
14. *Дюличева Ю.Ю* $DFBSA$-алгоритмы синтеза эмпирического леса «ссылкам» // Тезисы докладов Международной научной конференции «On Problems of Decision Making and Control under Uncertainties». – 2003. – С.95-97.
15. *Дюличева Ю.Ю.* Принятие решений на основе индуктивной модели ємпирического леса // Искусственый интеллект. – 2002. –№ 2. – С.110-115.
16. *Донской В.И., Дюличева Ю.Ю* Алгоритмы синтеза r-редуцированного эмпирического леса // Матемачиеские методы распознавания образов (ММРО-11): Тез.докл. – 2003. – С.71-74
17. *Донской В.И.,Башта А.И.* Дискретные модели принятия решения при неполной информации – Симферополь: Таврия, – 1992. – 166с.