

Витяев Е.Е.¹,
Костин В.С.,
Туровцева И.В.

Система «естественной» классификации с «прозрачным» интерфейсом

В предыдущих работах была определена «естественная» классификация, удовлетворяющая основным требованиям естествоиспытателей. В данной работе мы покажем, что «естественная» классификация может быть сделана «прозрачной» для пользователя. Прозрачность означает, что пользователь может не знать ничего о работе метода – все этапы работы метода могут быть продемонстрированы пользователю на объектах его выборки. В данной работе приводится описание системы «естественной» классификации, и её «прозрачного» интерфейса.

Ключевые слова: Интеллектуальный анализ данных, «естественная» классификация, Data Mining, классификация, кластерный анализ.

Введение

В предыдущих работах [Витяев, 1983, 1992, 2005; Витяев, Костин, 2009] была определена «естественная» классификация, удовлетворяющая требованиям естествоиспытателей. В этих работах было показано, что эти требования могут быть выведены из следующего принципа: «Разбиение объектов на классы должно производиться в соответствии с закономерностями, которым удовлетворяют объекты. Точнее, объекты одного класса должны подчиняться одной группе закономерностей, а объекты разных классов – разным группам закономерностей. Объекты одного класса также должны обладать некоторой целостностью. Целостность – взаимная согласованность закономерностей каждой группы по взаимному предсказанию свойств объектов» [Витяев, 1983]. В работах [Борисова, Загоруйко, 2004; Zagoruiko, Borisova, 2005] «естественная» классификация строится на основании специальной меры близости объектов. В работах [Демин, Витяев, 2010; Vityaev, Kostin, et al 2002; Vityaev, et al, 2008] приведен алгоритм и результаты эксперимента по построению «естественной» классификации для нуклеотидных последовательностей ДНК.

Как показано в работе [Витяев, Костин, 2009], понятие «естественной» классификации также тесно связано с понятием онтологии предметной области. Если к онтологии добавить множество закономерностей, обнаруживающих взаимосвязь признаков между собой, то мы можем построить «естественную» классификацию и систематику объектов предметной области. Онтология специфицирует *предмет исследования* - совокупность *существенных свойств (атрибутов)* и *отношений* объектов исследования, описываемых в некоторой *системе понятий* предметной области [Витяев, 2010]. Предмет исследования и онтология определяют «взгляд», «точку зрения», с которой рассматриваются (описываются в системе понятий) объекты предметной области, отношения и их свойства. Например, человек может рассматриваться с точки зрения разных систем понятий в таких областях, как археология, анатомия, медицина, психология и т.д. *Предметная область* – это совокупность *объектов предметной области*, рассматриваемых с точки зрения некоторого *предмета исследования*. Как показано в работе [Витяев, 2010] предметная область может быть задана эмпирической системой $\mathfrak{S} = \langle A, \Omega \rangle$, где A – множество объектов предметной области, а Ω – система понятий, а также множество отношений и операций величин, интерпретируемых в системе понятий предметной области.

«Естественная» классификация в силу приведённого выше определения, интерпретируема в онтологии предметной области, т.к. множество закономерностей, описывающих классы, интерпретируемы в системе понятий предметной области.

В данной работе мы покажем, что «естественная» классификация не только интерпретируема в онтологии предметной области, но и может быть сделана *прозрачной* для пользователя.

¹ Работа выполнена при финансовой поддержке гранта РФФИ № 11-07-00560-а, интеграционными проектами СО РАН № 47, 115, 119, а также Советом по грантам Президента РФ и государственной поддержке ведущих научных школ, проект НШ-3606.2010.1

2. Свойство прозрачности

Прозрачность «естественной» классификация означает, что пользователь может не знать ничего о работе метода – все этапы работы метода могут быть продемонстрированы пользователю на объектах его выборки.

Кроме того, прозрачность позволяет решить проблему согласования субъективных экспертных знаний с результатами объективного анализа данных, полученных в результате «естественной» классификации объектов.

Проблема согласования субъективных экспертных знаний с результатами объективного анализа состоит в том, что субъективное экспертное знание может следующим образом соотноситься с объективными закономерностями, классами и совокупностями признаков класса, полученных в процессе «естественной» классификации:

1. обнаруженные объективные закономерности, классы и признаки, в основном, соответствуют экспертному знанию, и тогда эксперт может дополнительно проверить:
 - 1.1. что объективные знания, полученные по выборке, полностью подтверждают субъективное экспертное знание;
 - 1.2. если обнаруженная закономерность, класс или совокупность признаков содержит меньше признаков, чем должно быть в соответствии с экспертным знанием, то эксперт может решить, что хотя правило и совместимо с его знанием, но оно не достаточно надежно. Разработанный интерфейс позволяет эксперту для обнаруженной закономерности, класса или совокупности признаков посмотреть конкретно на каких объектах они обнаружены. Тогда возможно два варианта согласования:
 - 1.2.1. эксперт признает, что на этих объектах нужны с его точки зрения признаки есть, тогда можно сделать вывод, что данных не достаточно для статистически значимого включения нужных признаков и надо увеличить выборку и заново пересчитать данные;
 - 1.2.2. эксперт видит, что в реальных данных нет тех признаков, которые с его точки зрения должны быть, тогда он должен сделать вывод о том, что его знания не соответствуют реальным данным. Тогда возможно два варианта:
 - 1.2.2.1. эксперт преувеличивает значимость и необходимость каких-либо признаков и ему надо пересмотреть своё знание и выяснить, почему его знания не соответствуют действительности. Опыт экспертов по знаниям показывает, что сформированные знания эксперта могут иметь следующие источники:
 - 1.2.2.1.1. большой пакет теоретических знаний, полученных в процессе обучения;
 - 1.2.2.1.2. статьи, обзоры, и другая литература, которые не имеют прямого отношения к его данным;
 - 1.2.2.1.3. мнения учителей и частные мнения других исследователей, у которых был свой собственный опыт.Поэтому эксперту необходимо прояснить для себя источник его знания и соотнести этот источник с его данными и решить, имеет ли его знание отношение к его данным;
 - 1.2.2.2. либо выборка данных собрана односторонне и в неё не включены данные с нужными признаками. В этом случае надо заново получить выборку и пересчитать данные;
 - 1.3. если обнаруженная на реальных данных закономерность, класс или совокупность признаков содержат признаки, которых не должно быть, с точки зрения эксперта, тогда эксперт также должен сделать вывод о том, что его знания не соответствуют реальным данным. Разработанный интерфейс в этом случае также позволяет пользователю для обнаруженной закономерности, класса или совокупности признаков класса посмотреть объекты, на которых они обнаружены. После анализа объектов эксперт может сделать два заключения:
 - 1.3.1. эксперт незнаком или недооценивает значимость некоторых признаков и ему надо пересмотреть своё мнение и выяснить, анализируя источники своих знаний в соответствии со случаями 1.2.2.1.1 - 1.2.2.1.3, почему его знания не учитывают данные признаки;

- 1.3.2. либо решить, что выборка данных собрана односторонне и поэтому в неё включены данные с обнаруженными признаками. В этом случае надо заново получить выборку и пересчитать данные;
2. обнаружены закономерности, классы или совокупности признаков, которые не знакомы эксперту. В этом случае эксперт также должен посмотреть, используя интерфейс, на каких объектах проявляются эти закономерности, классы и совокупности признаков и убедиться, что они действительно имеют место на этих данных. Тогда получим два случая:
 - 2.1. эксперт убеждается, что обнаруженная закономерность, класс или совокупность признаков действительно верны и тогда система обогащает знания эксперта;
 - 2.2. в результате анализа объектов, на которых обнаружена закономерность, класс или синдром, эксперт может решить, что данные собраны односторонне и рассматриваемые случаи надо расширить и заново пересчитать данные;
3. обнаруженные закономерности, классы или совокупности признаков противоречат его знанию. В этом случае эксперт также должен посмотреть, на каких объектах проявляются эти закономерности, классы и совокупности признаков и убедиться, что они действительно имеют место на его данных. После этого ему надо рассмотреть два случая:
 - 3.1. Эксперт может признать, что его знания, имеющие источники 1.2.2.1.1 - 1.2.2.1.3, не имеют под собой реальных оснований для его данных и не применимы к ним. Тогда система обогащает опыт эксперта;
 - 3.2. закономерности, классы или совокупности признаков были обнаружены на односторонне или тенденциозно полученных данных, которые требуют пересмотра². Обучающие данные должны быть собраны заново и пересчитаны.

В результате проделанной работы эксперт действительно получит новые знания о предметной области, т.к. «Знания – это воспринятая, осознанная и ставшая личностно значимой информация» [Бешенков, Ракитина, 2002].

3. Требования к системе.

В результате проделанного анализа были сформулированы следующие требования к системе:

1. понятный и удобный интерфейс, позволяющий пользователю легко общаться с системой;
2. гибкую настройку параметров – установку значений условной вероятности и пороговых значений критериев;
3. возможность обрабатывать большие массивы данных, для чего использовать СУБД SQLite, что позволяет не требовать сторонних серверов БД;
4. обеспечить кроссплатформенность с целью широкого применения системы, для чего осуществить её реализацию на языке Java.

4. Описание работы и интерфейса системы

Простота интерфейса была заимствована из широко распространенного в своё время интерфейса Norton Commander, только выполненного на современном уровне и содержащего не два окна, а три. Рабочее пространство системы (см. рис. 1) состоит из трёх окон, в которые загружаются данные (объекты) и их признаки, закономерности и совокупности признаков (синдромы, идеализированные описания).

Первое окно (два левых столбца) предназначены для отображения объектов и их атрибутов (рис. 1). Загрузить объекты можно используя соответствующую команду меню «Загрузить список объектов».

² В математической статистике уделяется большое внимание процедурам, обеспечивающим репрезентативность формируемой выборки. Репрезентативность можно определить как свойство выборочной совокупности представлять параметры генеральной совокупности, значимые с точки зрения задач исследования. Это связано с тем, что подход математической статистики основан на использовании ограниченного по объёму массива данных – выборочной совокупности (описаний эмпирических объектов) для характеристики такого теоретического объекта, как генеральная совокупность. Получив оценки (с некоторыми случайными погрешностями) параметров генеральной совокупности, исследователь может применять это знание для выводов относительно вновь поступающего эмпирического материала.

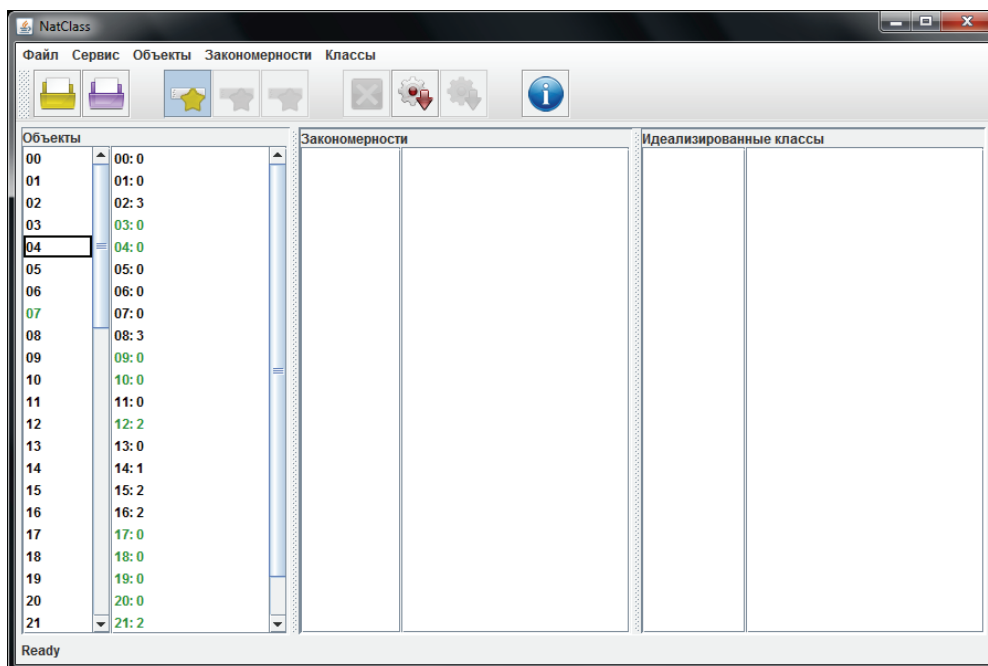


Рис. 1. Первое окно - объекты.

Загруженные объекты можно сравнить между собой по их атрибутам, используя функцию «Анализ объектов» на панели инструментов. Например, можно выбрать один объект (07) из списка объектов, а затем, нажав «Анализ объектов» (см. рис. 1), можно выбрать другой объект (04) и тогда в списке его атрибутов зеленым будут подсвечены атрибуты, совпадающие с атрибутами объекта (07). Тем самым всегда можно узнать, чем похожи между собой объекты.

После загрузки объектов в системе можно работать с закономерностями. Их можно загрузить, если они уже, хотя бы частично, были обнаружены, используя пункт меню «Загрузить список закономерностей». После чего в центральной части рабочего пространства появится список закономерностей с атрибутами (см. рис 2). Либо закономерности можно обнаружить на загруженных объектах. Для обнаружения закономерностей нужно задать следующие параметры: порог условной вероятности закономерности, уровень значимости критерия Фишера, длину закономерностей с полным перебором, а также возможность сохранять все полученные цепочки или только терминальные.

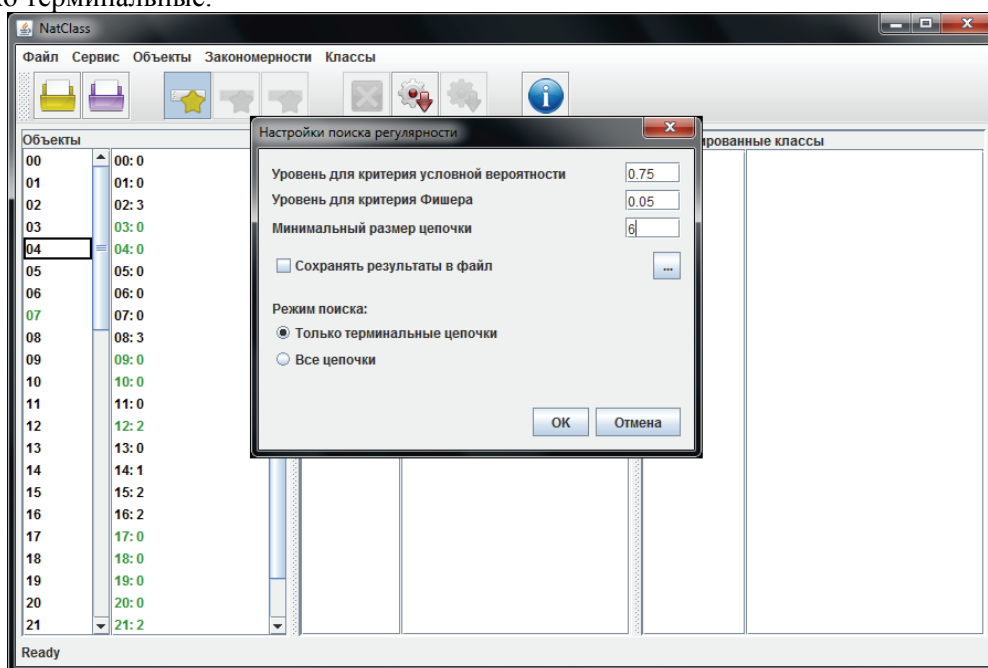


Рис. 2. Второе окно - закономерности.

После запуска алгоритма обнаружения закономерностей, список закономерностей будет постоянно пополняться. В нижней строке окна будет отображаться процент выполнения и

затраченное время на работу.

После того, как все закономерности будут обнаружены, либо алгоритм будет остановлен принудительно (в меню есть функция «Старт/Стоп» для остановки и продолжения работы), можно приступить к анализу полученных закономерностей, выбрав одну из них и нажав «Анализ закономерностей».

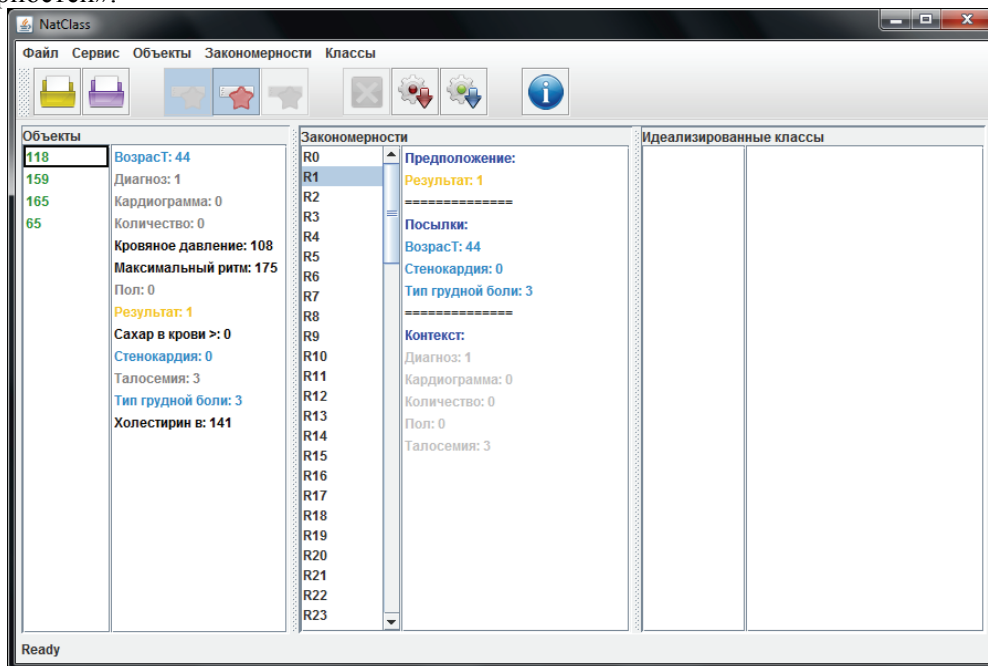


Рис. 3

Например, выбрав из полученного списка одну закономерность (R1) и начав анализ, можно проанализировать, для каких объектов эта закономерность выполняется (см. рис 3). При этом у выбранного объекта (118) голубым цветом будут выделены признаки посылки закономерности, а желтым - предсказываемый признак, как и в самой закономерности.

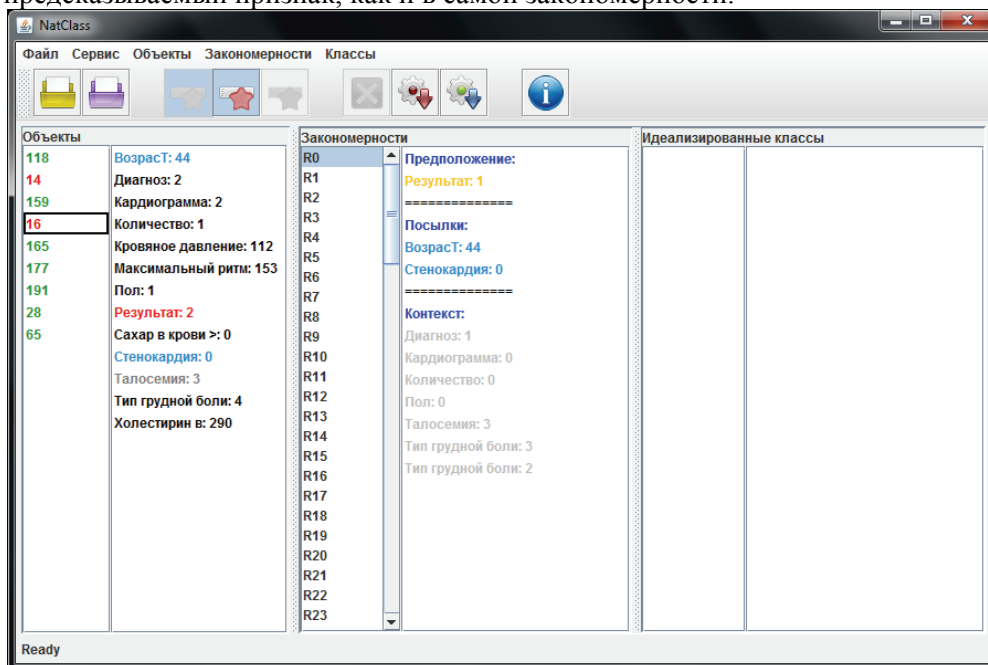


Рис. 4.

В случае, когда посылка закономерности истинна на объекте, а предсказываемый предикат ложен (опровержение закономерности на данном объекте), то такие объекты выделяются красным цветом (см. рис 4). Если выделить этот объект, например 16, как показано на рис. 4, то признаки посылки будут также подсвечены синим, а предсказываемый признак будет подсвечен красным.

Когда и объекты, и закономерности уже загружены в систему, становится доступным поиск классов и совокупностей признаков их описаний (синдромов, идеализированных классов). Так

же, как и для закономерностей, уже обнаруженные (хотя бы частично) классы можно загрузить, или обнаружить (см. рис. 5).

Для обнаружения классов надо зайти в пункт меню «Поиск идеальных классов». После завершения работы алгоритма или после загрузки, в третьем окне появится список классов, построенный по имеющимся закономерностям.

Система позволяет проанализировать полученные классы и совокупности признаков следующим образом:

1. для каждого класса и его описания можно проследить, какой набор закономерностей на нем выполняется в соответствии с принципом «естественной» классификации – «разбиение объектов на классы должно – производиться в соответствии с закономерностями, которым удовлетворяют объекты класса». Например, если взять класс (C169) (см. рис. 6) и осуществить анализ этого класса, нажав пункт меню «Анализ класса», то во втором окне отобразятся закономерности, характеризующие этот класс, а в первом окне – объекты, относящиеся к данному классу;

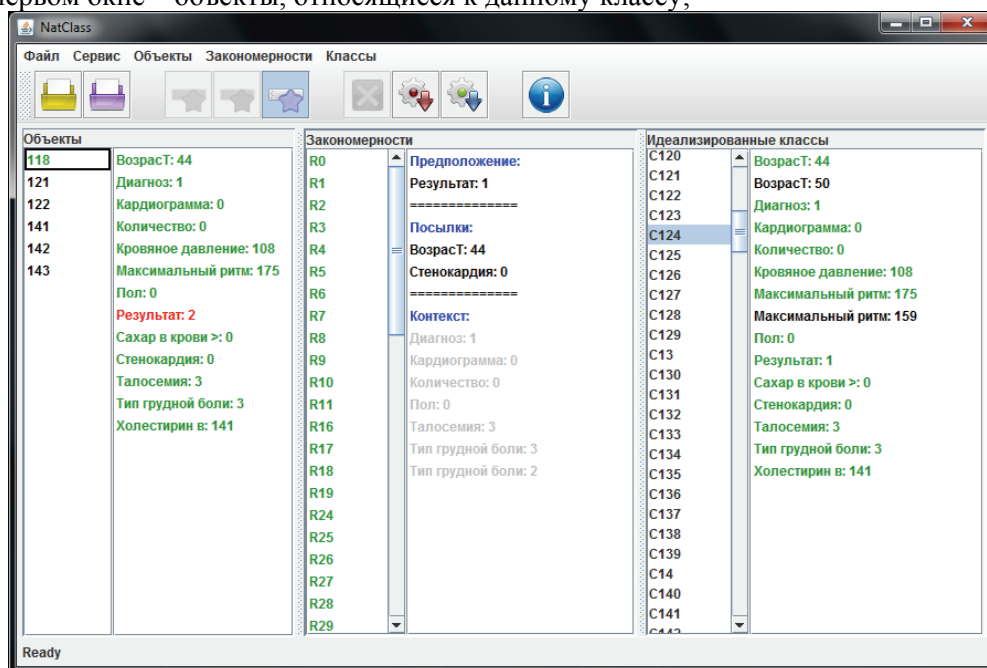


Рис. 5. Идеализированные классы.

2. если во втором окне выделить некоторую закономерность, например R53, характеризующую класс C169, то признаки закономерности R53 высветятся в классе C169 тем же цветом, что и в закономерности, показывая нам эти признаки в признаках описания класса. Это позволяет в совокупности признаков класса (синдроме, идеализированном описании класса) увидеть, как эти признаки взаимно предсказывают друг друга, образуя структурный закон строения объекта [Витяев, Костин, 2009]. Переводя курсор с закономерности на закономерность, можно увидеть всю структуру взаимосвязей признаков класса;
3. в первом окне при этом отображаются объекты, принадлежащие выбранному классу, например для класса C124 это объекты 118, 121, 122, 141-143 (см. рис. 5). При этом на объекте показываются зеленым признаки объекта, вошедшие в совокупность признаков описания класса, а красным, не вошедшие в это описание. В признаках класса зеленым будут подсвечены признаки объекта, вошедшие в описание класса, а черным, добавленные в процессе построения класса. Всё это позволяет:
 1. увидеть в описании объекта признаки (отмеченные зеленым), которые вошли в совокупность признаков класса, и, значит, существенные для принадлежности к данному классу – информативные признаки объекта;
 2. увидеть в описании объекта признаки, отмеченные красным, которые не вошли в описание класса, и, значит, не информативные. Кроме того, если этот признак есть и в описании объекта, и в описании класса, как, например, «результат» на рис. 6, но имеет разные значения, то это означает возможную ошибку или искажение данных;

3. в совокупности признаков класса показываются существенные признаки выделенного объекта. Чёрным отмечены те признаки класса, которые были добавлены к признакам объекта в процессе классификации и, значит, необходимые с точки зрения других значений признаков класса, т.к. они предсказываются по закономерностям класса;
4. переводя курсор с объекта на объект, можно в совокупности объектов класса посмотреть, как складывается описание класса в совокупность признаков класса (синдром) и почему этот синдром характерен для этих объектов.

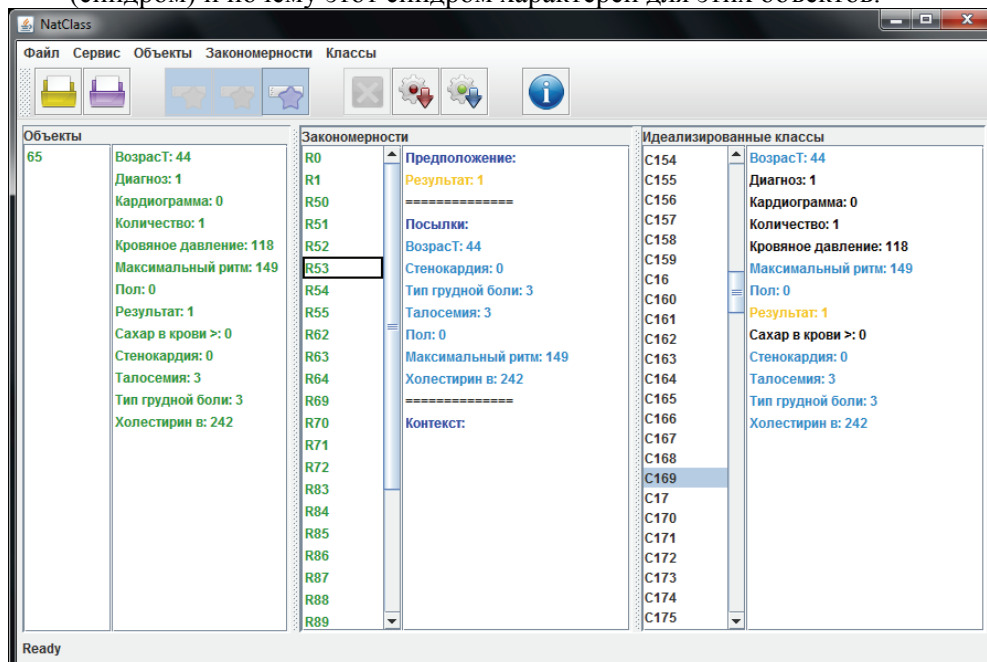


Рис. 6

Система позволяет в любой промежуточной стадии получить общие сведения о состоянии рабочей среды: количество загруженных объектов и их атрибутов, количество обнаруженных закономерностей и уровень условной вероятности и критерия Фишера на них, а также количество полученных классов. Все закономерности, классы и их описания сохраняются в файлы базы данных SQLite и могут быть впоследствии использованы, как для анализа, так и в других целях.

ЛИТЕРАТУРА

- Бешенков С.А., Ракитина Е.А. Моделирование и формализация. Методическое пособие. – М. // Лаборатория Базовых Знаний, 2002: 336с.
- Витяев Е.Е. Классификация как выделение групп объектов, удовлетворяющих разным множествам согласованных закономерностей // Анализ разнотипных данных (Вычислительные системы вып. 99), Новосибирск, 1983: с. 44-50.
- Витяев Е.Е. Естественная классификация как закон природы // Интеллектуальные системы и методология. ("Материалы научно-практического симпозиума "Интеллектуальная поддержка деятельности в сложных предметных областях", Новосибирск – 7-9 апреля 1992) // Новосибирск – 1992. Вып. 4.
- Витяев Е.Е. Естественная классификация и систематика как законы природы // Анализ структурных закономерностей (Вычислительные системы Вып. 174), Новосибирск – 2005.
- Витяев Е.Е., Костин В.С. Естественная классификация, систематика, онтология. Информационные технологии в гуманитарных исследованиях, Вып. 13, ИАЭТ СО РАН, Новосибирск, 2009: с. 65-75
- Витяев Е.Е. Извлечение информации из данных // Информационные технологии в гуманитарных исследованиях, Вып. 15, ИАЭТ СО РАН, Новосибирск, 2010: с. 9-16.
- Демин А.В., Витяев Е.Е. Метод построения «естественной» классификации // Информационные технологии в гуманитарных исследованиях, Вып. 15, ИАЭТ СО РАН, Новосибирск, 2010: с. 16-22
- Борисова И.А., Загоруйко Н.Г. "Естественная классификация" // Сборник трудов ИАИ-2004, Киев, 2004: с. 33-42.
- Материалы сайта <http://www.math.nsc.ru/AP/ScientificDiscovery>
- Vityaev E.E., Kostin V.V., Podkolodny N.A., Kolchanov N.A. NATURAL CLASSIFICATION OF NUCLEOTIDE SEQUENCES. // Proc. of the Third International Conference On Bioinformatics of Genome Regulation and Structure (BGRS'2002, Novosibirsk, Russia, July 14-20, 2002), v3, ICG, Novosibirsk, 2002: p. 197-199
- Vityaev E.E., Lapardin K.A., Khomicheva I.V., Proskura A., L. Transcription factor binding site recognition by regularity matrices based on the natural classification method. Intelligent Data Analysis. Special issue: "New Methods in Bioinformatics Presented at the fifth International Conference on Bioinformatics of Genom Regulation and Structure" eds. Evgenii Vityaev and Nikolai Kolchanov. v.12(5), IOS Press, 2008: p. 495-512
- Zagoruiko N., Borisova I. "Principles of natural classification"// Pattern Recognition and Image Analysis, 2005, Vol.15, No.1: p.27-29.