

УДК 621.396.961.1



МЕТОДЫ DATA MINING И OLAP ДЛЯ РЕШЕНИЯ ЗАДАЧ БАНКОВСКОГО КРЕДИТОВАНИЯ

О.В. Ивченко¹, А.С. Коряк², В.Б. Репка³, А.В. Шерстнюк⁴

¹ ХНУРЭ, г. Харьков, Украина, ivchenko.o@mail.ru

² ХНУРЭ, г. Харьков, Украина, akoryak@mail.ru

³ ХНУРЭ, г. Харьков, Украина, victoria_repka@kture.kharkov.ua

⁴ ХНУРЭ, г. Харьков, Украина, troll_andrey@mail.ru

Рассмотрены проблемы, возникающие в процессе кредитования физических лиц в банках Украины. Определены задачи, которые необходимо решить для эффективного кредитования и минимизации риска. Предлагается использовать нейросетевую систему поддержки принятия решений, основанную на методах Data Mining (модульная нейронная сеть) и OLAP-технологии.

МОДУЛЬНАЯ НЕЙРОННАЯ СЕТЬ, РЕКУРРЕНТНЫЙ МНОГОСЛОЙНЫЙ ПЕРСЕПТРОН, СИСТЕМА ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ, БАНК, КРЕДИТОВАНИЕ, OLAP, DATA MINING, БАЗА ЗНАНИЙ, ХРАНИЛИЩЕ ДАННЫХ

Введение

Главная цель банка как коммерческого предприятия – получение прибыли. Основным источником прибыли для банка – предоставление в ссуду своих кредитных ресурсов. В то же время основные убытки банки терпят от непосредственно кредитной деятельности. Причиной такого положения является проведение некоторыми банками слишком рискованной кредитной политики. Естественно, полностью избежать риска невозможно, поскольку предоставление кредитов – это изначально рискованный вид бизнеса. Поэтому одной из главных задач банка является минимизация кредитного риска. Эта деятельность позволит банкам максимизировать прибыль и снизить потери от проведения кредитных операций.

Существуют различные подходы и методы оценки кредитоспособности частного заемщика, основанные на статических моделях или базирующиеся на субъективных оценках специалистов банка и зависящие от личного опыта и впечатления о потенциальном клиенте, а также автоматизированные системы, созданные с использованием математических моделей.

Наиболее популярными сегодня в области кредитования являются скоринговые системы, основанные на технологии углубленного анализа данных Data Mining, которые предоставляют банковскому служащему данные в удобном для принятия решений виде и способствуют выработке рекомендаций.

Искусственные нейронные сети и другие методы Data Mining давно используются большинством зарубежных банков не только для определения кредитоспособности заемщика, но и для многих других финансовых задач и за это время успели проявить себя с лучшей стороны и принести немалый доход.

Постановка задачи. На основании анализа процесса кредитования в украинской банковской системе предлагается разработать нейросетевую систему поддержки принятия решений (НССППР), в которой в качестве механизма выработки решений на основании хранящейся или поступающей информации используется модульная нейронная сеть.

В статье описывается структура данной системы, которая способствует повышению точности прогнозирования при выдаче кредита физическим лицам и снижению риска кредитования в банках. НССППР основана на методах Data Mining (модульной нейронной сети) и OLAP-технологии.

1. Проблемы кредитования физических лиц в банках

Банк является центральным звеном в системе рыночных отношений, развитие его деятельности – необходимое условие реального создания рыночной экономики. Банк – своего рода посредник между элементами экономической системы. Аккумулируя финансовые ресурсы у одних экономических агентов (имеющих их излишек) и распределяя их другим (имеющим недостаток финансовых ресурсов), банк выполняет важнейшую роль в народном хозяйстве – роль источника средств для развития и роста перспективных секторов экономики.

Любой банк, взаимодействуя с внешней средой как открытая система, находится в динамической связи с другими объектами народного хозяйства. К последним, прежде всего, относятся предприятия корпоративного сектора – юридические лица и индивидуальные предприниматели.

Но на сегодняшний день для банков Украины наиболее актуальной является проблема кредитования физических лиц. Одной из ключевых проблем является получение объективной, достоверной

и полной информации о физическом лице. Банк сталкивается с непрозрачностью расходов и доходов клиента, отсутствием истории взаимоотношений с клиентом и отсутствием гарантий доходов клиента на срок кредитования. На Западе эта проблема решается следующим образом: кредитное бюро предоставляет информацию о клиенте в режиме on-line, включая его кредитную историю, финансовое состояние и репутацию. Данной информации вполне достаточно для принятия решения о предоставлении или не предоставлении кредита. Причем в данном случае клиент не тратит время на сбор документации. У нас в стране порой для получения более или менее крупного кредита бухгалтерия вынуждена в течение месяца, а то и более заниматься подготовкой запрашиваемой документации, необходимой банку для принятия решения о предоставлении кредита.

В настоящее время в банках обрабатываются большие объёмы информации, но главной проблемой является не обработка информации, а формирование исходных данных для принятия решений. Это относится не только к банку, но и к любой другой организации, в которой циркулируют большие потоки информации. Но в банке задача обработки большого количества информации и принятия

правильного решения играет наиболее важную роль из-за сжатости сроков.

Большинство банков имеет филиалы и отделения и, как правило, использует несколько различных информационных систем, что приводит к возникновению проблемы разрозненного хранения данных.

2. Структура нейросетевой системы поддержки принятия решений

Для комплексного подхода к решению задачи кредитования физических лиц в банке предлагается создать единую банковскую систему управления процессом кредитования с единым хранилищем данных, форматом данных, разработанную на единой платформе, содержащей новейшие технологии обработки данных и методы принятия решений – нейросетевую систему поддержки принятия решений (СППР).

Предлагается общая структура данной системы, представленная на рис. 1.

Система содержит базу знаний, модуль настройки и обучения НС, блок принятия решений и интерфейсы эксперта и пользователя.

База знаний состоит из хранилища данных, в которое посредством OLAP системы поступают дан-

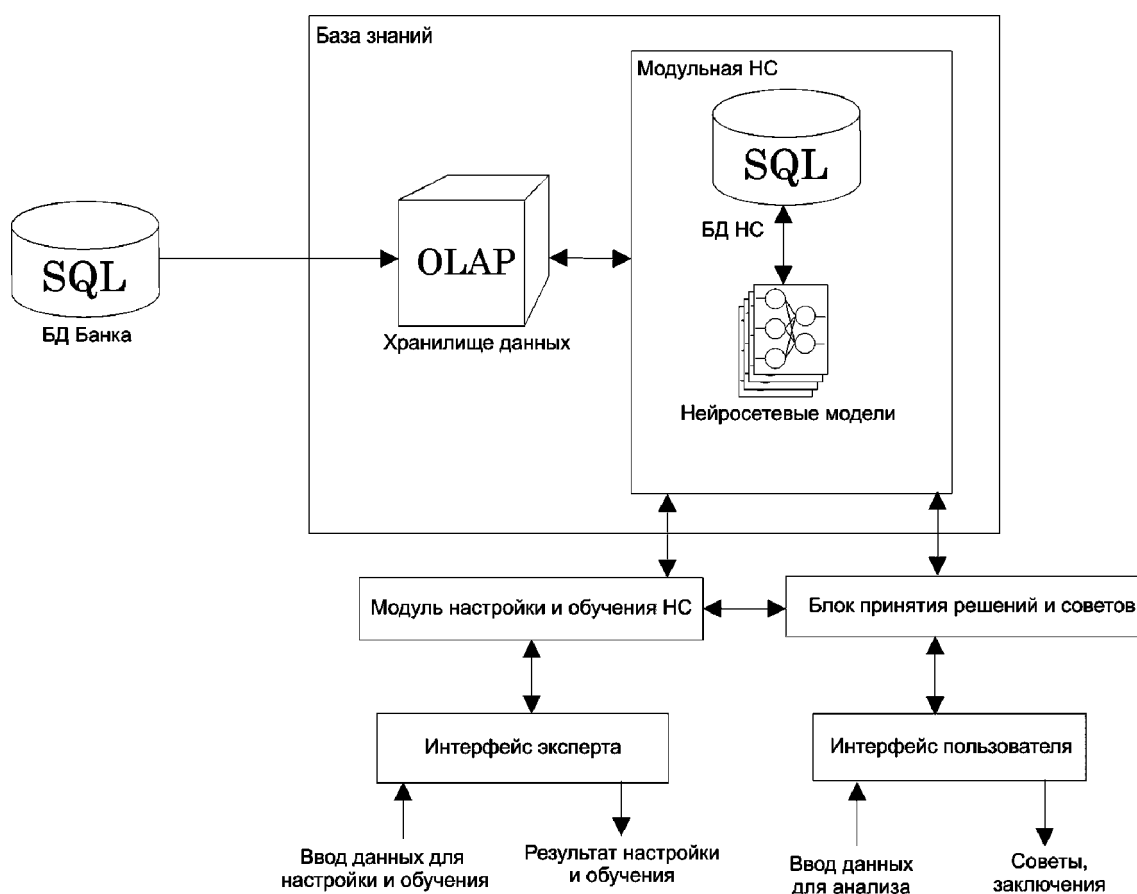


Рис. 1. Структура нейросетевой системы поддержки принятия решений для банковского кредитования физических лиц

ные из баз данных всех филиалов банка и модульной нейронной сети.

Модульная нейронная сеть содержит нейросетевые модели — изначально рекуррентные, но в будущем и другие НС и базу данных, хранящую данные для обучения НС.

Построение моделей НС производит эксперт, который подаёт данные для обучения НС через интерфейс эксперта; данные поступают в «Модуль обучения и настройки НС», после чего НС обучаются на основании знаний эксперта и характеристик предшествующих заемщиков, поступивших из «Хранилища данных». Настройки для нейросетевых моделей заносятся в базу данных НС.

Конечным пользователем программного продукта является сотрудник отдела кредитования, к которому клиент обращается с просьбой о выдаче ему кредита. Сотрудник банка вводит через интерфейс пользователя информацию о новом заемщике, затем эта информация попадает в «Блок принятия решений и советов», откуда поступает на модульную НС. В модульной НС поставленную задачу решает та НС, входящая в её состав, которая выдает наиболее точное решение.

Нейронная сеть, настроенная опытным экспертом и обученная на предыдущем опыте, принимает решение и подаёт его на блок принятия решений, там принятое решение оценивается и в качестве одного из возможных правильных решений предоставляется пользователю или, в случае, если данных о подобных ситуациях в обучающей выборке нет — передается в «Модуль настройки и обучения НС». На основании полученного совета кредитный эксперт располагает обоснованными данными с целью выработки дальнейших решений — выдать кредит или отказать, а полученные результаты и информация о рассмотренной заявке и заемщике поступает в хранилище данных. По истечении определенного промежутка времени НС могут корректироваться в связи с ростом объемов информации, что позволяет предоставить руководству банка качественную информацию для принятия решения о выдаче кредита.

Использование системы аналитической обработки данных в режиме реального времени — OLAP (On-Line Analysis Processing) позволяет выбирать данные одновременно из нескольких источников, структурировать их и формировать информацию, необходимую для принятия правильных и экономически обоснованных решений.

В НССППР (рис. 1) данные поступают из базы данных, содержащей первичную информацию, далее информация проходит через следующие уровни хранения и анализа: хранилище данных (ХД), OLAP.

Сбор первичной информации представляет собой накопление подробной информации в базе данных за прошлые годы банковской деятельности. Базы данных содержатся как в банке, так и в его филиалах. Данные источники информации являются разрозненными и могут использовать различные СУБД, так как программное обеспечение в филиалах может обновляться гораздо позже, чем в самом банке. Для СППР важным критерием является количество и качество первичной информации, поэтому требуется собрать, структурировать и представить в удобном виде для последующих этапов анализа и принятия решений всю имеющуюся информацию.

Необходимая для анализа информация поступает в хранилище данных. Хранилище данных — предметно-ориентированный, интегрированный, неизменяемый, поддерживающий хронологию набор данных, организованный для целей поддержки принятия решений. Концепция ХД предполагает разделение структур хранения данных для оперативной обработки и выполнения аналитических запросов. Это позволяет в рамках одной СППР объединить две (и более) подсистемы [1].

Существует виртуальное и физическое ХД. В данном случае используется физическое ХД. Данные из БД всех филиалов и самого банка переносятся в единое хранилище, к которому адресуются все аналитические запросы.

При переносе данных из БД в ХД наиболее важной задачей является их очистка. Очищенные данные сохраняются и могут использоваться для анализа и принятия на их основе решений.

Существуют определённые классы задач, которые лучше решать в рамках хранилища данных, в нашем случае это анализ клиентской базы. Анализ клиентской базы позволяет сформировать целевые сегменты клиентов и использовать эту информацию при кредитовании. Целевые сегменты формируются на основе демографических и финансовых показателей, а также других параметров клиентов.

3. Модульная нейронная сеть как механизм выработки решений

Использование методов интеллектуального анализа данных, в частности нейронных сетей (НС), для решения подобных задач вызвано их нелинейностью, нетребовательностью к глубокому пониманию связей между исходными данными и результатами в отличие от традиционных методов. Важнейшее свойство НС, свидетельствующее об их огромном потенциале и широких прикладных возможностях, состоит в параллельной обработке информации одновременно всеми нейронами. Благодаря этой способности при большом количестве межнейрон-

ных связей достигается значительное ускорение обработки информации. Не менее важное свойство НС состоит в способности к обучению и обобщению полученных знаний.

Ядром НСППР является модульная нейронная сеть [2]. Структура сети представлена на рис. 2.

Нейронная сеть состоит из модулей RNN 1 – RNN N, представляющих собой рекуррентные сети и блок принятия решений в виде однослойного перцептрона SLP. Такая структура является частным случаем общей структуры модели модульной сети [3]. В качестве модулей используются различные архитектуры рекуррентных нейронных сетей. Рекуррентные нейронные сети были выбраны как наиболее подходящие для решения задачи классификации динамических объектов [4], которыми являются сведения о физических лицах при кредитовании. Возможными архитектурами, которые составляют модульную НС, являются сети:

- RMLP (рекуррентный многослойный перцептрон);
- RTRN Вильямса Зипсера;
- сеть Эльмана.

На данный момент программно реализованы нейронная сеть RMLP и RTRN.

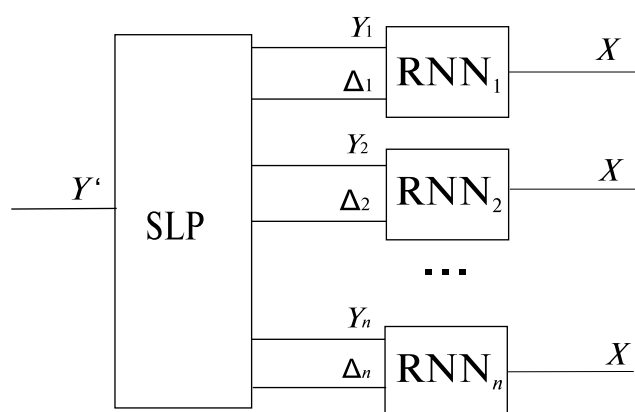


Рис. 2. Структура модульной НС

В соответствии с предложенной структурой модульной НС разработан алгоритм обучения, состоящий из следующих шагов:

Обучение модулей RNNi.

1. Подача на все выходы модулей RNNi вектора X, содержащего характеристики клиента банка.
2. Вектор X обрабатывается каждым i-м модулем, в результате на выходе модулей получаем n выходных векторов Yi.
3. Используя выходы Yi и вектор идеальных выходов D, модули RNNi обучаются по соответствующим алгоритмам обучения.
4. После полного обучения модулей получаем векторы Delta i, представляющие собой множество оценок качества обучения модулей RNNi:

а) суммарная квадратическая ошибка:

$$E = \sqrt{\frac{1}{SN_o} \sum_S \sum_j (d_j^S - y_j^S)^2}, \quad (1)$$

где S – количество элементов в обучающем множестве; N_o – количество нейронов в выходном слое; d_j^S – известное решение S-го элемента из обучающего множества; y_j^S – решение, полученное нейронной сетью.

б) средняя относительная ошибка:

$$\sigma = \frac{1}{SN_o} \sum_S \sum_j \left(\frac{|d_j^S - y_j^S| + 1}{|d_j^S| + 1} - 1 \right) \cdot 100\% \quad (2)$$

в) процент правильно распознанных образов:

$$\eta := \frac{S}{S_0} \cdot 100\%, \quad (3)$$

где S₀ – количество образов в обучающем множестве; S – количество правильно спрогнозированных образов.

Обучение блока принятия решений.

1. Подача на входы всех модулей RNNi вектора X, содержащего характеристики клиента банка из обучающей выборки.

2. Вектор X обрабатывается каждым модулем, в результате на выходе модулей получаем набор выходных векторов Yi.

3. На вход сети SLP (Single layer perceptron) подается входной вектор, состоящий из векторов Yi и оценок качества обучения нейронных сетей Delta i.

4. После обработки входного вектора на выходе сети SLP получаем вектор Y', представляющий собой набор нулей и единиц и содержащий двоичный код индекса i модуля RNNi (номер модуля в двоичной системе счисления), у которого наименьшая среднеквадратическая ошибка на выходе. На основании этой ошибки обучается сеть SLP.

Алгоритм работы обученной модульной сети заключается в следующем:

1. Подача на все выходы модулей RNNi вектора X, содержащего характеристики нового клиента банка. Вектор X обрабатывается каждым модулем, в результате на выходе модулей получаем n выходных векторов Yi.
2. На вход сети SLP подается входной вектор, состоящий из векторов Yi и оценок качества обучения нейронных сетей Delta i. После обработки входного вектора на выходе сети SLP получаем вектор Y', представляющий собой двоичный код индекса i модуля RNNi, у которого наименьшая среднеквадратическая ошибка на выходе.
3. На выходе модульной нейронной сети получаем вектор Yk, где k – индекс модуля с наименьшей среднеквадратической ошибкой, определенной сетью SLP.

Приведем результаты экспериментов по использованию в качестве модулей рекуррентный многослойный перцептрон RMLP (Recurrent MultiLayer Perceptron), который является динамической сетью и характеризуется запаздыванием входных и выходных сигналов, объединяемых во входной вектор сети. Структура сети RMLP, содержащая один скрытый слой, один входной и один выходной сигнал, представлена на рис. 3.

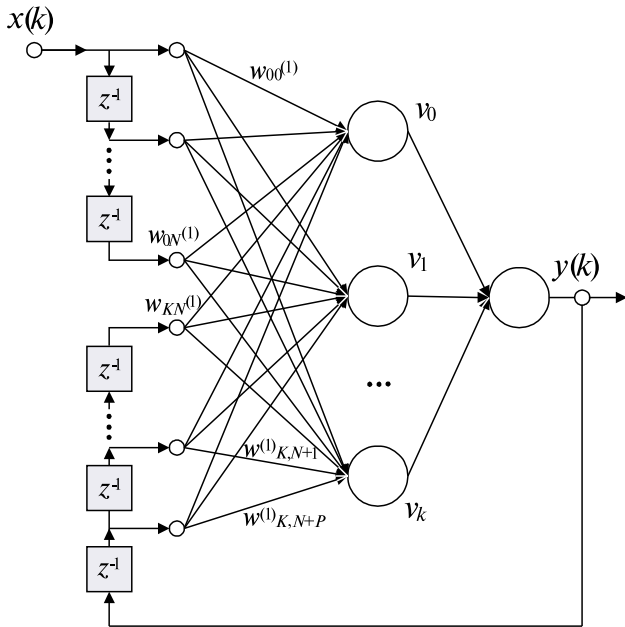


Рис. 3. Структура рекуррентного многослойного перцептрона

Такая система реализует отображение:

$$\begin{aligned} y(k+1) &= f(x(k), x(k-1), \dots \\ & x(k-(N-1)), y(k-1), \\ & y(k-2), \dots, y(k-P)), \end{aligned} \quad (4)$$

где $(N-1)$ – количество задержек входного сигнала; P – количество задержек выходного сигнала. Обозначим K количество нейронов в скрытом слое. В этом случае RMLP можно характеризовать тройкой чисел (N, P, K) . Подаваемый на вход вектор сети имеет вид:

$$\begin{aligned} x(k) &= [x(k), x(k-1), \dots \\ & x(k-(N-1)), y(k-P), \\ & y(k-P+1), \dots, y(k-1)]^T. \end{aligned} \quad (5)$$

Все нейроны сети имеют сигмоидальную функцию активации.

Моделируемая сеть RMLP включает в себя три слоя: входной, один скрытый и выходной. Сеть адаптируется с применением градиентного алгоритма обучения [4]. Данный алгоритм функционирует в режиме «online», принимая поступающие входные данные и соответствующие им значения ожидаемого вектора d и оперативно корректируя значения весов.

Выходной вектор сети RMLP зависит от совокупности входных и выходных векторов в предыдущие моменты времени. Если на вход НС подавать образы в одной и той же последовательности, то на графике среднеквадратической ошибки сети RMLP (рис. 4), можно заметить резкие скачки значений среднеквадратической ошибки через каждые 150 итераций, которые обусловлены тем, что на вход сети образы начинают подаваться сначала. Если бы данные на вход сети поступали непрерывно, отображая какой-либо динамический процесс, то график среднеквадратической ошибки был бы гладким.

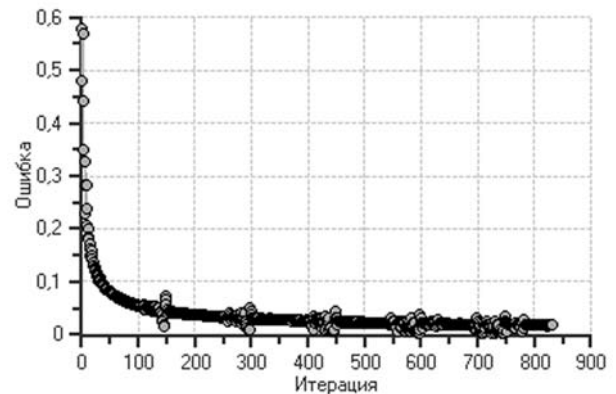


Рис. 4. График среднеквадратической ошибки сети RMLP

Для успешного применения сети RMLP для решения задачи кредитования физических лиц, то есть задачи классификации – определения, к какому классу можно отнести нового заемщика, к «вернувшему» или «не вернувшему» кредит, можно представить сам процесс кредитования как динамический и непрерывный.

В качестве запаздывающих входных и выходных сигналов берутся те элементы выборки, которые соответствуют заемщикам, получившим кредит в последнее время и характеристики которых отражают текущую экономическую ситуацию. Это позволяет ускорить обучение НС и получить результаты, отражающие состояние кредитующихся при существующем экономическом уровне страны.

На вход сети, с установленным количеством запаздывших входных и выходных сигналов, подаётся новая информация о заемщике, сеть при этом обучается непрерывно. В то время, как на динамическую сеть, такую как многослойный перцептрон, придётся переобучать через определённые промежутки времени, когда будут накапливаться большие объёмы новой информации, на что потребуется достаточно много времени.

Для сравнения качества и скорости обучения был реализован многослойный перцептрон. Результаты обучения показали, что НС RMLP адаптируется гораздо быстрее и при меньшем количестве скрытых

слоёв и нейронов в них. Например: если обучить данные нейронные сети на одну и ту же выборку (150 образцов, каждый входной вектор представляет собой 32 входных значения и одно значение требуемого выхода), то одинаковые значения среднеквадратических ошибок сетей получаются при обучении многослойного персептрона за 1000–1500 итераций, а рекуррентного многослойного персептрона за 500 итераций.

При изменении параметров для обучения нейронной сети, а также при изменении количества нейронов в скрытом слое, можно добиться лучших результатов, например, на рис. 5 изображён график среднеквадратической ошибки сети RMLP для той же выборки, но с другими значениями количества нейронов, параметра наклона функции активации и коэффициентами скорости обучения.

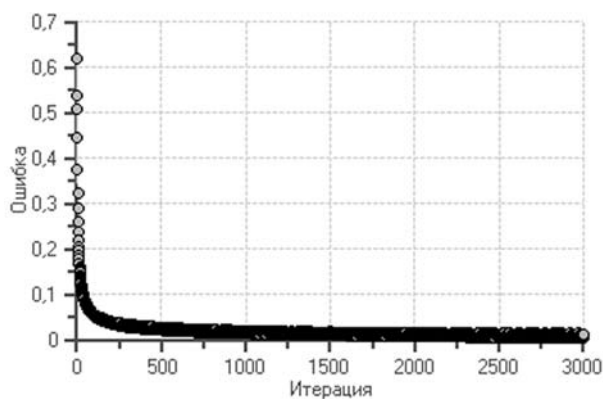


Рис. 5. График среднеквадратической ошибки сети RMLP с изменёнными параметрами

Таким образом, можно сделать вывод, что сеть RMLP обоснованно можно включить в качестве модуля предложенной модульной сети и использовать для решения задачи кредитования.

Заключение

В статье была предложена структура нейросетевой системы поддержки принятия решений (НССППР), в которой в качестве механизма выработки решений на основании хранящейся или поступающей информации используется модульная нейронная сеть.

Использование хранилища данных и OLAP позволит решить проблему разрозненного хранения данных в банках и сформировать исходные данные для принятия решений.

Внедрение предложенных нейросетевых моделей кредитования и создание на их основе СППР позволит обеспечить рациональное использование кредитных ресурсов банка и прогнозировать сроки возврата и процентные ставки таким образом, чтобы получить максимальную прибыль от кредитования.

Список литературы: 1. Барсегян А.А. Методы и модели анализа данных: OLAP и Data Mining. — СПб.: БХВ-Петербург. — 2004. — 336 с. 2. Bart L.M. Happel, Jacob M.J. Murre. Design and Evolution of Modular Neural Architectures. Neural Networks, Vol.7, No 6/7, 985-1004 pp. 1994. 3. Eric Ronko, Peter Gawthrop Modular Neural Networks: a state of the art / Technical report: CSC-95026, Centre for System and Control University of Glasgow, Glasgow, UK. — 1995. 22 pp. 4. Осовский С. Нейронные сети для обработки информации / Пер. с польского И.Д. Рудинского. — М.: Финансы и статистика, 2002. — 344 с.

Поступила в редколлегию 24.10.2007