Ключевые слова:
анализ финансового состояния, интеллектуальный анализ данных, методология анализа, прогнозирование показателей финансовой отчетности, методы Data Mining, метод опорных векторов, программное обеспечение

***Н. А. Никифорова**, к. э. н., проф., зав. кафедрой «Экономический анализ» Академии бюджета и казначейства Минфина России (**e-mail: oomka-nna@inbox.ru**)*
***Л. В. Донцова**, д. э. н., проф. кафедры управления капиталом Российской академии народного хозяйства и государственной службы при Президенте Российской Федерации (**e-mail: dontsova.L@mtu-net.ru**)*
***Е. В. Донцов**, инженер факультета вычислительной математики и кибернетики МГУ им. М. В. Ломоносова (**e-mail: don-and-home@mail.ru**)*

Интеллектуальный анализ данных в моделировании финансового состояния предприятий

Оценка финансового состояния предприятия — важнейшее условие принятия правильных управленческих решений. Она состоит из трех больших взаимосвязанных блоков: анализ финансового положения и деловой активности; анализ финансовых результатов; оценка возможных перспектив развития. Основная задача, решаемая в рамках этого исследования, — прогнозирование финансового состояния предприятия на основе данных, представленных в его финансовых отчетах и других документах. В работе использован качественно новый инструмент — метод интеллектуального анализа данных.

Управление современным бизнесом немыслимо без прогнозирования и анализа данных, который в зависимости от целей исследования можно разделить на следующие виды:

- **Информационно-поисковый** и **визуальный анализ**. В ходе такого анализа мы, не приобретая никаких новых знаний о предмете, получаем возможность рассмотреть его по частям и с разных точек зрения. Осуществляется это, как правило, путем четко сформулированного запроса к реляционной базе данных. Этот вид анализа лежит в области **детализированных данных**, никак их не обобщая.
- **Оперативно-аналитический анализ**, или **OLAP**. В OLAP данные агрегируются, предоставляя аналитику возможность получить любую степень обобщения в любом разрезе. В отличие от информационно-поискового анализа, здесь мы можем обнаружить различного рода закономерности в данных, которые иначе были бы не видны. OLAP вводит нас в сферу **обобщенных данных**.

- **Интеллектуальный анализ**, или **Data Mining**. Этот вид анализа направлен на выявление скрытых закономерностей в данных, например, повторяющихся шаблонов или кластеров. Иначе говоря, на его основе можно получить модели, позволяющие лучше понимать данные и предсказывать их поведение. Data Mining в действительности предполагает непосредственное **обнаружение знаний**.

Начиная с середины 1990-х гг. XX в. в информационной индустрии наблюдается рост интереса к технологиям анализа данных, основанным на технологиях систем поддержки принятия решений. За ними закрепился ставший уже привычным в англоязычной литературе термин «Data Mining», или «Knowledge Discovery». «Data Mining» не имеет однозначного перевода на русский язык («добыча данных», «извлечение информации» и т. д.), поэтому в большинстве случаев используется в оригинале. Наиболее удачным непрямым переводом считается термин «интеллектуальный анализ данных».

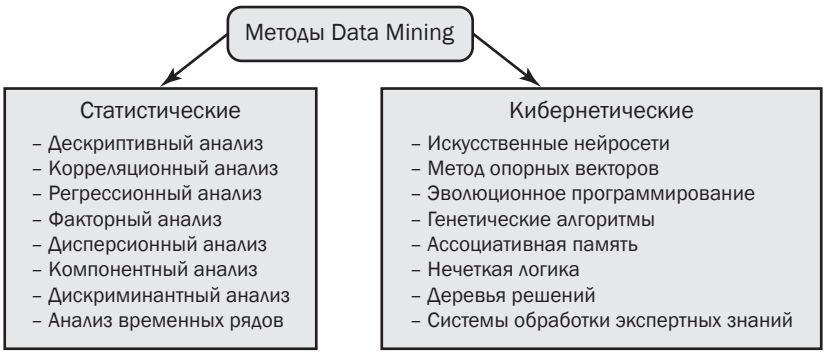
В 1960–1970-е гг. XX в. советские математики под руководством В. Н. Вапника разработали метод обобщенного портрета, основанный на построении оптимальной разделяющей гиперплоскости. Требование оптимальности заключалось в том, что обучающие объекты должны быть удалены от разделяющей поверхности настолько далеко, насколько это возможно. В 1990-е гг. метод получил мировую известность и после некоторой переработки и серии обобщений стал называться методом опорных векторов (support vector machines — SVM). Классическое определение этого термина предложено в 1996 г. в работе ученых У. Файада, Г. Пятецки-Шапиро, П. Смита «Нетривиальный процесс обнаружения новых, потенциально полезных, корректных и интерпретируемых закономерностей в данных».

Популярность Data Mining сегодня можно сравнить с популярностью этого направления полстолетия назад, на заре компьютерной эпохи. Тогда, правда, этот термин не был известен, но много говорили об искусственном интеллекте, о нейронных сетях и распознавании образов. Однако, за немногими исключениями, практическую реализацию теории пришлось отложить до тех пор, пока аппаратная и программная инфраструктура не развилась до современного уровня. И сегодня, по завершении пятидесятилетнего цикла развития, мы вновь обращаемся к решению задач анализа, уже обладая для этого мощными вычислительными системами и системами управления базами данных, развитой операционной и языковой средой.

Структура методов Data Mining представлена на рис. 1.

Рисунок 1

Методы интеллектуального анализа данных



Источник: составлено авторами.

Методы Data Mining делятся на две большие группы: **Supervised learning** (Обучение с учителем) и **Unsupervised Learning** (Обучение без учителя).

В первом случае задача анализа данных, например классификация, осуществляется в несколько этапов. Это один из способов машинного обучения, в ходе которого испытуемая система принудительно обучается с помощью примеров. Между входами и эталонными выходами может существовать некоторая зависимость, но она неизвестна. Известна только конечная совокупность прецедентов, называемая **обучающей выборкой**. На основе этих данных требуется восстановить зависимость, т. е. построить алгоритм, способный для любого объекта выдать достаточно точный ответ. Для измерения точности ответов, как и в обучении на примерах, может вводиться функционал качества. Сначала с помощью какого-либо алгоритма Data Mining строится модель анализируемых данных — классификатор. Затем классификатор подвергается обучению. Другими словами, проверяется качество его работы и, если оно неудовлетворительно, проводится дополнительное обучение классификатора. Так происходит до тех пор, пока мы не достигнем требуемого уровня качества или не убедимся, что выбранный алгоритм не подходит для работы с данными либо же сами данные не имеют структуры, которую можно выявить.

Обучение без учителя — способ машинного обучения, с помощью которого испытуемая система учится выполнять поставленную задачу спонтанно, без вмешательства со стороны экспериментатора. Как правило, это пригодно только для задач, в которых известны описания множества объектов (обучающей выборки) и требуется обнаружить внутренние взаимосвязи, зависимости, закономерности, существующие между объектами. Очевидно, что если эти закономерности есть, то модель должна их представить, и неуместно говорить об ее обучении. Отсюда и название данного способа.

Сфера применения методов интеллектуального анализа достаточно разнообразна. Приведем лишь некоторые направления:

- **Телекоммуникационный бизнес**. Телекоммуникационные компании работают в условиях жесткой конкуренции. Использование технологий Data Mining, направленных как на анализ доходности и риска клиентов, так и на защиту от мошенничества, может сэкономить этим компаниям огромные средства.
- **Промышленное производство** — идеальная среда для применения технологий Data Mining. Причина этого заключается в природе технологического процесса, который должен быть воспроизводимым и контролируемым. Таким образом создается статистическая стабильность, которая так важна для классификации. Пример применения Data Mining в промышленности — прогнозирование качества изделия в зависимости от параметров технологического процесса.
- **Банковский сектор**. Классическим примером применения Data Mining на практике может быть решение вопроса о кредитоспособности клиентов банка. Система поддержки принятия решений со встроенной функциональностью Data Mining опирается в своей работе только на базу данных банка, где записывается детальная информация о каждом клиенте и, в конечном итоге, факт его кредитоспособности. Классификационные алгоритмы Data Mining обрабатывают эти данные, и полученные результаты используются для принятия решений.
- **Страхование**. В этой сфере, так же, как в банковском деле и маркетинге, возникает задача обработки больших объемов информации для определения типичных групп (профилей) клиентов. Эта информация используется для того, чтобы предлагать определенные услуги страхования с наименьшим для компании риском и, возможно, с пользой для клиента.

Таким образом, применение методов интеллектуального анализа открывает новые перспективы во многих сферах исследования и в том числе, как мы попытаемся доказать, в области **анализа финансового состояния предприятий**.

В данной работе поставлена цель — использовать методы Data Mining в оценке финансового состояния предприятий Москвы. Количественные значения показателей для исследования были взяты непосредственно из финансовой отчетности 70 государственных унитарных предприятий и 80 акционерных обществ за 2006–2008 гг. Для решения задачи был разработан программный комплекс, а также проведена предварительная обработка показателей финансово-хозяйственной деятельности предприятий.

Задача прогнозирования сформулирована следующим образом. Дано некоторое количество k коэффициентов x за определенный период i:

$$x_{i1}, x_{i2}, ..., x_{ik}. \quad (1)$$

Необходимо найти функцию, переводящую данный вектор в число, характеризующее класс банкротства предприятия за следующий отчетный период i:

$$y_{i+1} = f(x_{i1}, x_{i2}, ..., x_{ik}). \quad (2)$$

Для решения поставленной задачи использовались методы интеллектуального анализа: метод нейросетей и метод опорных векторов (SVM).

Рассмотрим SVM как более эффективный способ классификации. Метод нейросетевого анализа находит лишь один и далеко не оптимальный способ разделения классов из всех возможных, а SVM заключается в построении разделяющей поверхности, наиболее удаленной от всех разделяемых точек. Таким образом, можно предположить, что качество распознавания новых примеров у SVM должно быть выше, чем у нейронной сети. Критерий останова для обучения нейронной сети — нулевая ошибка на обучающем множестве, а для метода опорных векторов — близость построенной разделяющей гиперплоскости к оптимальной.

Основное отличие SVM от метода нейросетей заключается в том, что для нейросети количество настраиваемых коэффициентов должно априорно задаваться пользователем на основании некоторых эвристических соображений. В методе опорных векторов количество параметров автоматически определяется во время настройки, и обычно оно меньше, чем число векторов в обучающей последовательности. Ненулевыми остаются коэффициенты опорных векторов, с помощью которых строится разделяющая гиперплоскость.

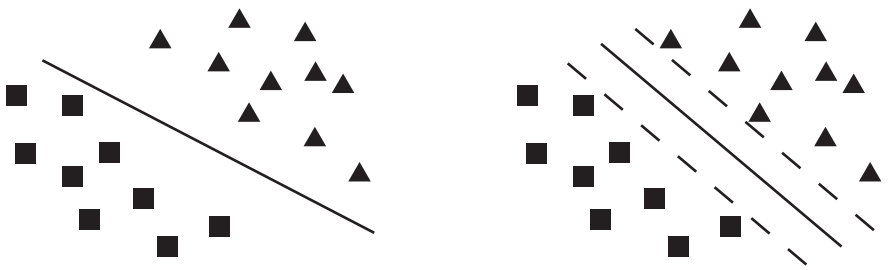
Метод опорных векторов позволяет получить функцию классификации с минимальной верхней оценкой ожидаемого риска (уровнем ошибки классификации). Он также дает возможность использовать линейный классификатор для работы с нелинейно разделяемыми данными. Его недостатком является неустойчивость по отношению к шуму в исходных данных. Шумовые выбросы обучающей выборки будут существенным образом учтены при построении разделяющей гиперплоскости.

МЕТОД ОПОРНЫХ ВЕКТОРОВ. ЛИНЕЙНО РАЗДЕЛИМЫЙ СЛУЧАЙ

Основной идеей SVM является то, что он позволяет находить линейное разделение двух множеств таким образом, чтобы расстояние между этими множествами и гиперплоскостью было максимальным. Такая гиперплоскость называется **оптимальной разделяющей**, в отличие от гиперплоскости, получаемой с помощью нейронных сетей и обеспечивающей лишь разделение классов, без оценки расстояния между множествами.

Рисунок 2

Разделение множеств нейросетью и SVM



Нейронная сеть | Метод опорных векторов

Представим, что имеется так называемое обучающее множество векторов, состоящее из N точек данных:

$$\{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_N, y_N)\},$$

где $\vec{x}_i \in R^n$ — известные параметры;
$y_i \in \{-1, +1\}$ — известные значения, определяющие принадлежность вектора к тому или иному множеству.

Допустим, множества линейно разделимы. Целью ставится поиск такой линейной разделяющей гиперплоскости H, называемой классификатором, чтобы

$$f(\vec{x}) = \text{sgn}((\vec{w}\cdot\vec{x}) - b),\ \vec{w} \in R^n,\ b \in R, \tag{3}$$

где $\vec{w}$ и b — параметры классификатора, подлежащие определению.

Для того чтобы построить гиперплоскость H, рассмотрим две параллельные гиперплоскости H_1 и H_2:

$$H_1: y = (\vec{w}\cdot\vec{x}) - b = +1; \tag{4}$$

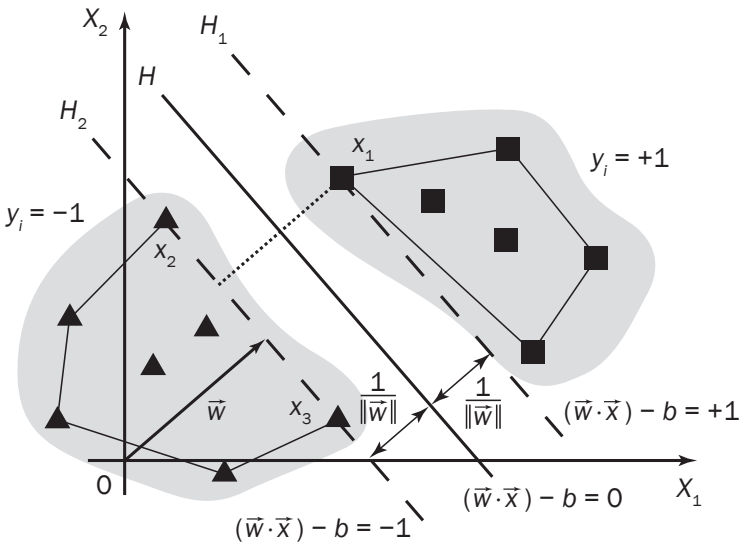
$$H_2: y = (\vec{w}\cdot\vec{x}) - b = -1, \tag{5}$$

с условиями максимизации расстояния между H_1 и H_2 и отсутствия точек данных между ними. Тогда H_1 содержит как минимум одну точку с $y = +1$, а H_2 — как минимум одну точку с $y = -1$. Расстояние между множествами определяется как расстояние между этими гиперплоскостями. Гиперплоскость H строится как параллельная равноудаленная от H_1 и H_2. Разделение в двухмерном пространстве представлено на рис. 3.

Точки, через которые проходят гиперплоскости H_1 и H_2, называются опорными векторами. Метод опорных векторов находит их линейную комбинацию, с помощью которой строится разделяющая гиперплоскость. Это точки x_1, x_2, x_3.

Рисунок 3

Построение гиперплоскости H. Линейное разделение



Примечание: жирная прямая соответствует гиперплоскости H — классификатору, построенному так, чтобы максимизировать расстояние между множествами треугольников и квадратов. По осям X_1 и X_2 отложены координаты точек разделяемых множеств.

Расстояние от H_1 до параллельной гиперплоскости H равно:

$$\frac{|(\vec{w}\cdot\vec{x}) - b|}{\|w\|} = \frac{1}{\|w\|}. \tag{6}$$

Расстояние между H_1 и H_2 равно $\frac{2}{\|w\|}$. Чтобы максимизировать это расстояние, минимизируется $\|w\|$ с условиями, что между H_1 и H_2 нет точек данных:

$$(\vec{w}\cdot\vec{x}) - b \geq +1 \text{ для точек из первого класса } y = +1, \tag{7}$$

$$(\vec{w}\cdot\vec{x}) - b \leq -1 \text{ для точек из второго класса } y = -1.$$

Неравенства (7) можно записать в виде

$$y_i((\vec{w}\cdot\vec{x}) - b) \geq 1 \tag{8}$$

и решать задачу:

$$\min_{\vec{w},\, b} \frac{1}{2}\vec{w}^T\vec{w} \tag{9}$$

с линейными ограничениями (8). Таким образом, задача свелась к хорошо известной задаче квадратичного программирования, которая решается минимизацией функционала Лагранжа:

$$L(\vec{w}, b, \vec{\alpha}) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \alpha_i (y_i((\vec{w}\cdot\vec{x}) - b) - 1) \to \min, \quad (10)$$

где $\alpha_i \geq 0$ — множители Лагранжа.

Так как Лагранжиан является выпуклой функцией, и минимум ищется по выпуклому множеству, можно свести поиск к эквивалентной двойственной задаче — максимизации L по α, исходя из того, что производные L по $\vec{w}$ и b равны нулю:

$$\frac{\partial L}{\partial \vec{w}} = 0\ ,\ \frac{\partial L}{\partial b} = 0\ ,\ \text{откуда следует, что } \vec{w} = \sum_{i=1}^{N} \alpha_i y_i \vec{x}_i \ \text{ и } \ \sum_{i=1}^{N} \alpha_i y_i = 0.$$

Двойственная задача формулируется следующим образом: максимизировать

$$L_D(\vec{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (\vec{x_i}, \vec{x_j}) \quad (11)$$

с ограничениями $\alpha_i \geq 0$, $\sum_{i=1}^{N} \alpha_i y_i = 0$, $i = 1, \ldots, N$. (12)

Значение b определяется из условия Куна-Таккера:

$$\alpha_i [y_i((\vec{w}\cdot\vec{x}_i) - b) - 1] = 0,\ i = 1, \ldots, N. \quad (13)$$

Решения не существует для линейно неразделимых данных, т. к. в этом случае не существует гиперплоскостей H_1 и H_2.

Для того чтобы, несмотря на это, построить разделение классов, нужно ослабить неравенства (7), т. е. ввести набор небольших положительных чисел $\xi_i \geq 0$, $i = 1, \ldots, N$. Тогда (7) примет вид:

$$\vec{w}\cdot\vec{x}_i + b \geq 1 - \xi_i \ \text{ для } y_i = 1 \quad (14)$$

$$\text{и } \vec{w}\cdot\vec{x}_i + b \leq -1 + \xi_i \ \text{ для } y_i = -1\ .$$

В таком случае следует минимизировать не $\|w\| = \vec{w}^T\vec{w}$, а $\frac{1}{2}\vec{w}^T\vec{w} + C\sum_i \xi_i$ при условии, что $y_i(\vec{w}\cdot\vec{x}_i + b) + \xi_i \geq 1$, где $\xi_i \geq 0$.

Лагранжиан будет иметь следующий вид:

$$L(\vec{w}, b, \xi, \alpha, \mu) = \frac{1}{2}(\vec{w}, \vec{w}) + C\sum_{i=1}^{N} \xi_i - \sum_{i=1}^{l} \alpha_i \{[(\vec{x}_i\cdot\vec{w}) + b]y_i + \xi_i - 1\} - \sum_{i=1}^{N} \mu_i \xi_i.$$

Проводя рассуждения, аналогичные линейно разделимому случаю, получим, что требуется максимизировать функцию:

$$\max_{\vec{\alpha}} L_D(\vec{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (\vec{x}_i, \vec{x}_j) \quad (15)$$

с условиями $0 \le \alpha_i \le C$, $\sum_{i=1}^{N} \alpha_i y_i = 0$, $i = 1, ..., N$. (16)

Эта задача отличается от задачи (11) – (12) только условиями ограниченности сверху множителей Лагранжа. Мы выбирали C таким образом, чтобы минимизировать ошибку классификации.

Весовые коэффициенты $\vec{w}$ разделяющей гиперплоскости могут быть определены из формулы:

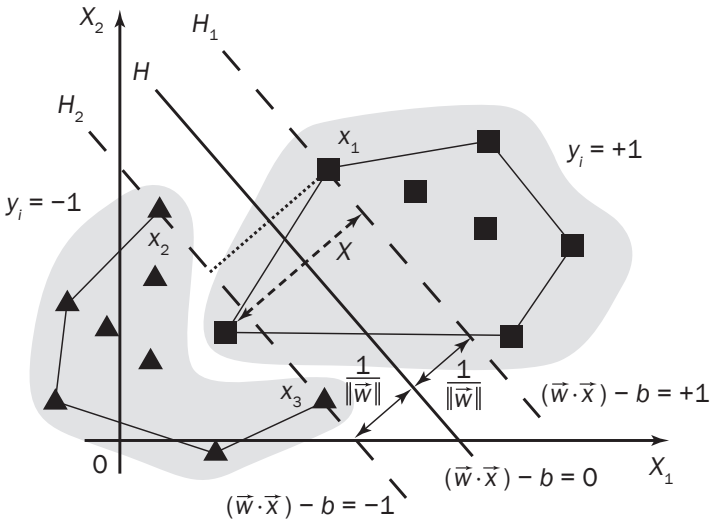
$$\vec{w} = \sum_{i=1}^{N} \alpha_i y_i \vec{x}_i. \quad (17)$$

ЛИНЕЙНО НЕРАЗДЕЛИМЫЙ СЛУЧАЙ. НЕЛИНЕЙНОЕ РАЗДЕЛЕНИЕ

Чаще всего бывает, что классы линейно неразделимы.

Рисунок 4

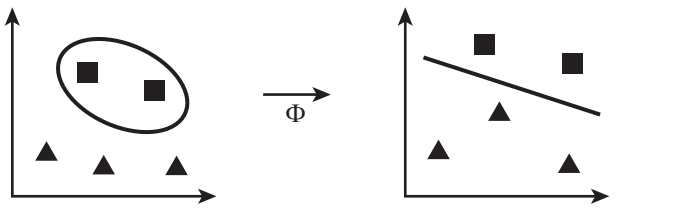
Пример линейно неразделимого случая



Тогда требуется найти нелинейное разделение. Еще в 1974 г. В. Вапником был описан способ построения такого разделения. Нужно перевести входные векторы в пространство более высокой размерности, в котором они будут линейно разделимы.

Рисунок 5

Перевод входного вектора в другое пространство



Это преобразование $\vec{x}_i \to \Phi(\vec{x}_i)$. Тогда в выражении (14) $(\vec{x}_i, \vec{x}_j) \to \Phi(\vec{x}_i) \cdot \Phi(\vec{x}_j) = K(\vec{x}_i, \vec{x}_j)$, где K называют ядром. Выбор функции ядра зависит от специфики решаемой задачи. Наиболее часто используемые ядра — полиномиальное $K(\vec{x}, \vec{y}) = (\vec{x} \cdot \vec{y} + 1)^p$, различные радиальные $K(\vec{x}, \vec{y}) = e^{-\rho \cdot d(\vec{x}, \vec{y})}$. Особенно популярно Гауссово ядро при $d(\vec{x}, \vec{y}) = \|\vec{x} - \vec{y}\|_2^2$, $\rho = \frac{1}{2\sigma^2}$.

Итак, в общем случае задача классификации изображений свелась к задаче поиска максимума квадратичного функционала с ограничениями, т. е. к задаче квадратичного программирования:

$$L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j}^{N} \alpha_i \alpha_j y_i y_j K(\vec{x}_i, \vec{x}_j) \xrightarrow{\alpha} \max, \tag{18}$$

$$0 \le \alpha_i \le C,$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0.$$

Функцией решения будет являться:

$$f(x) = \operatorname{sgn}\left(\sum_{\substack{support \\ vectors}} \alpha_i y_i K(\vec{x}_i \vec{x}) + b \right),$$

где сумма берется только по опорным векторам, т. е. таким векторам, что $\alpha_i \neq 0$. Этот факт важен для SVM, потому что обычно число опорных векторов меньше, чем число обучающих.

Для решения задачи нахождения разделяющей гиперплоскости большую роль играют условия Куна-Таккера. Для нашей задачи необходимые и достаточные условия максимума:

$$\alpha_i \{[(\vec{x}_i \cdot \vec{w}) + b] y_i + \xi_i - 1\} = 0, \tag{19}$$

$$\mu_i \xi_i = 0.$$

Используя описанный алгоритм поиска α, мы будем проверять, удовлетворяет ли текущее α условиям Куна-Таккера. Нарушение этих условий происходит в случае, если

$$0 < \alpha_i < C, \tag{20}$$

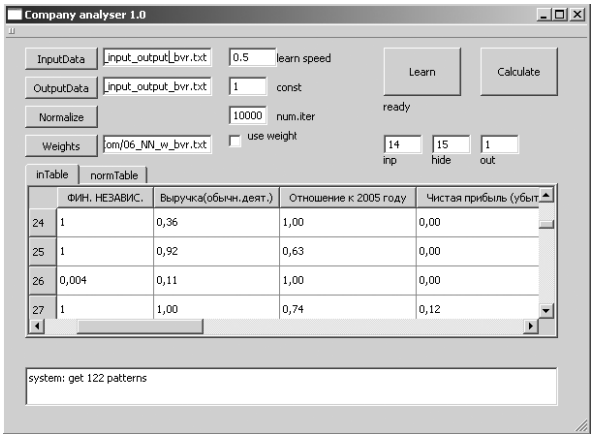
а $R_i = y_i [(\vec{x}_i \cdot \vec{w}) + b - y_i] \neq 0$.

РАСЧЕТ И ОЦЕНКА ВЕРОЯТНОСТИ БАНКРОТСТВА

Для решения задачи прогнозирования финансового состояния предприятия была использована программа А. А. Лукьяницы TestSVM, реализующая поиск опорных векторов. Также для преобразования данных из экономических отчетов написан программный комплекс Company analyzer на языке C++ с использованием библиотеки QT. Данное расширение позволяет компилировать программы как в ОС Windows, так и в Linux без изменений в исходном коде. Программа обрабатывает таблицы отчетов о финансово-хозяйственном состоянии предприятий, проводит предварительную подготовку данных для использования методов Data Mining, а также осуществляет прогнозирование.

Рисунок 6

Внешний вид программы Company analyzer



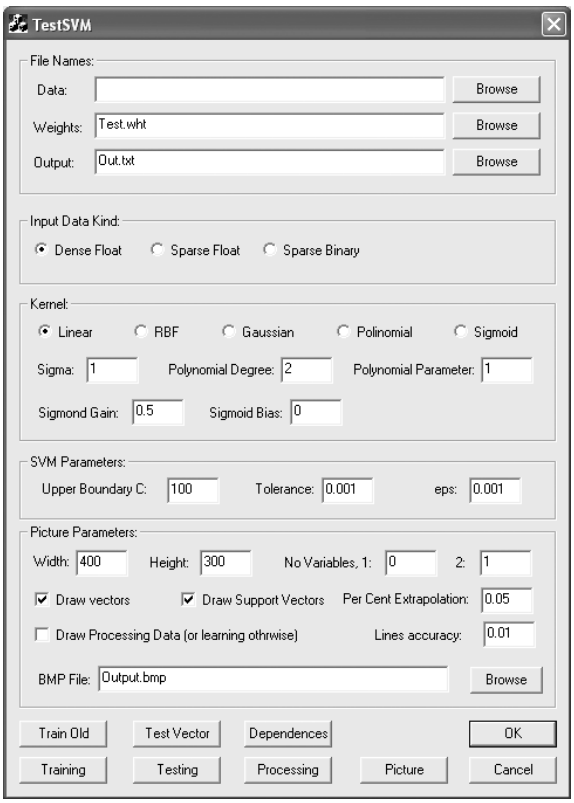
Программа Company analyzer работает с входными файлами текстового формата, включающего показатели для оценки предприятия. Подготовка данных к исследованию проводится в два этапа. На первом этапе необходимо выбрать исходные данные по каждому предприятию в две таблицы — за 2006 и 2007 гг. Данные первой таблицы за 2006 г. преобразуются в расчетные значения от 0 до 1 (нормализуются), второй таблицы за 2007 г. — распределяются на классы банкротства в зависимости от выбранного метода классификации.

Полученные файлы используются для поиска опорных векторов. Предприятия, содержащиеся в них, являются «обучающим множеством». Каждая строка первого файла эквивалентна одному предприятию, а числа в ней являются преобразованными данными этого предприятия. В каждой строке второго файла находится число — индикатор принадлежности предприятия к одному из классов банкротства в 2007 г. Таким образом, количество подобных файлов будет равно количеству классов выбранного метода классификации.

Вторым подготовительным этапом является поиск опорных векторов. Для этого полученные данные обрабатываются с помощью программы TestSVM. В качестве входных данных указывается таблица, содержащая данные 2006 г., в качестве выходных — принадлежность каждого предприятия к одному из классов банкротства в 2007 г.

Рисунок 7

Внешний вид программы TestSVM



Результатом работы программы является файл, содержащий коэффициенты опорных векторов. Данный файл позволит спрогнозировать, будет ли предприятие принадлежать к определенному классу банкротства в 2008 г. Подготовка к прогнозированию на этом завершена. Далее осуществляется сам процесс прогнозирования. Он заключается в том, что нормализованные данные предприятий 2007 г. загружаются в программу TestSVM. Далее необходимо загрузить файл, содержащий коэффициенты опорных векторов определенного класса банкротства. В выходном файле программы будет столбец чисел, показывающих, принадлежит или не принадлежит предприятие к выбранному классу банкротства в 2008 г.

Чтобы проверить правильность прогноза с помощью метода опорных векторов, необходимо сравнить классы банкротства, рассчитанные выбранным методом, с прогнозируемым опорными векторами значением.

Исследование проводилось на основании информации о финансово-хозяйственной деятельности 122 предприятий Москвы за 2006–2008 гг. В работе использованы три подхода к классификации предприятий: методика, используемая собственником предприятий Москвы (СПМ), метод Донцовой – Никифоровой и метод Бивера. В процессе работы был проведен поиск опорных векторов с применением линейного разделения пространства, а также радиальной базисной функции Гаусса.

Во множестве предприятий по методике, используемой СПМ, получено 19 компаний с признаками банкротства, 103 — без его признаков. Классификация предприятий по методу Донцовой – Никифоровой следующая: 56 предприятий — I класса, 27 — II и III, 7 предприятий — IV и 5 — V класса. Классификация по методу Бивера: 102 предприятия — I класса, 3 предприятия — II и 17 — III класса.

Для каждого класса вычисляется гиперплоскость (опорные вектора), которая отделяет компании, относящиеся к данному классу, от всех остальных компаний. Применение линейного разделения пространства дало следующие результаты на обучающем множестве (табл. 1).

Таблица 1

Результаты поиска опорных векторов (линейное разделение пространства)

	Метод Бивера			Метод Донцовой – Никифоровой					Метод СПМ
Класс	I	II	III	I	II	III	IV	V	–
Ошибка классификации, %	12	0	4	26	22	21	3	1,6	0,8
Кол-во опорных векторов	49	11	17	82	89	58	21	15,0	57,0

Вычисленные опорные вектора могут быть использованы для прогнозирования. Для ответа на вопрос о корректности прогноза необходимо провести тестирование. Данными для тестов служит информация о предприятиях за 2008 г. Мы должны на основе данных за 2007 г. спрогнозировать уже известные значения классов банкротства предприятий в следующем, 2008 г.

Тестовое множество состояло из 17 компаний, классифицированных по методике СПМ, и 45 предприятий, классифицированных по методам Бивера и Донцовой – Никифоровой. Процентное совпадение прогноза метода опорных векторов с обучающим множеством метода опорных векторов показано в табл. 2.

Таблица 2

Совпадение прогноза метода опорных векторов с обучающим множеством, %

Метод Бивера			Метод Донцовой – Никифоровой					Метод СПМ
I	II	III	I	II	III	IV	V	–
80,4	91,3	78,0	63,0	65,2	73,9	89,1	91,3	82,4

Следует отметить неравномерное распределение предприятий по классам (метод Бивера): I классу соответствует 35 предприятий, II — 3, III — 7. Из этих предприятий корректно было спрогнозировано 33 (94,3 %) предприятия в I классе, 2 (66 %) — во II и 3 (43 %) — в III классе.

В тестируемом множестве распределение предприятий по классам методом Донцовой – Никифоровой также различно: I классу соответствует 15 предприятий, из них было распознано корректно 9 (60 %); II классу — 15 предприятий, III — 10, IV — 3 и V — 2 предприятия. Данные предприятия не были корректно спрогнозированы методом опорных векторов с линейным разделением пространства.

Тестирование метода прогнозирования, применяемого СПМ, удалось провести на 17 предприятиях. При этом совпадение с обучающей выборкой произошло в 14 случаях (82,4 %). Предприятий с признаками банкротства было 4, из них корректно спрогнозировано 1.

Известно, что линейное разделение пространства методом опорных векторов не является самым эффективным на множестве объектов со сложными границами. В связи с этим было проведено обучение и прогнозирование методом опорных векторов с радиальной базисной функцией Гаусса (табл. 3).

Таблица 3

Результаты поиска опорных векторов (радиальная базисная функция Гаусса)

	Метод Бивера			Метод Донцовой – Никифоровой					Метод СПМ
Класс	I	II	III	I	II	III	IV	V	–
Ошибка классификации, %	1,6	0	1	2	2	2	1	0	0,8
Кол-во опорных векторов	64,0	15	29	77	74	65	29	22	57,0

После нахождения гиперплоскостей был проведен прогноз состояния предприятий на 2008 г. по имеющимся данным за 2007 г. Процентное совпадение прогноза метода опорных векторов с обучающим множеством показано в табл. 4.

Таблица 4

Совпадение прогноза метода опорных векторов с обучающим множеством, %

Метод Бивера			Метод Донцовой – Никифоровой					Метод СПМ
I	II	III	I	II	III	IV	V	–
82,6	93,5	82,6	78,3	84,8	82,6	91,3	93,48	82,4

По методу Бивера I классу соответствует 35 предприятий, II — 3, III — 7. Из этих предприятий корректно было спрогнозировано 32 (91,4 %) предприятия в I классе, 2 (66 %) — во II и 4 (57 %) — в III классе.

Распределение предприятий по классам методом Донцовой – Никифоровой следующее: I классу соответствует 15 предприятий, из них было распознано корректно 14 (93,3 %); II — 15 предприятий, из них распознано 9 (60 %); III классу принадлежит 10, распознано 7 (70 %); IV — 3, распознано 0; V классу — 2 предприятия, распознано 1 (50 %).

Проведенное прогнозирование позволяет говорить об эффективности применения метода опорных векторов, когда пространство делится не линейным ядром, а радиальной базисной функцией Гаусса. Как видно из обучения, ошибка классификации обучающего множества при линейном разделении велика. Низкие значения обусловлены крайне небольшим количеством принадлежащих к классам предприятий. Малая выборка — главная причина низкого качества распознавания классов. Однако в тех классах, где количество предприятий достаточно велико, распознавание проходит эффективно. Из этого можно сделать вывод о необходимости увеличения

обучающей выборки. Если нельзя увеличить количество предприятий за конкретный отчетный период, то обучающую выборку можно увеличить за счет расширения временного промежутка данных о финансово-хозяйственной деятельности исследуемых предприятий.

Оценка возможных перспектив развития организации в настоящее время крайне важна. Анализ финансового состояния предприятия позволяет судить о степени эффективности вложений в него. Мы провели прогнозирование показателей финансовой отчетности на следующий отчетный период. Для этой цели был применен такой современный метод интеллектуального анализа данных, как метод опорных векторов. Сравнение полученного прогноза с истинным положением дел в этом периоде позволяет сделать вывод об эффективности прогнозирования. Результаты исследования: метод опорных векторов в среднем достигал корректности прогноза в 85 % случаев. Таким образом, прогнозирование можно считать успешным.

Библиография

1. Вапник, В. Н., Червоненкис, А. Я. Теория распознавания образов. — М.: Наука, 1974.
2. Донцова, Л. В., Никифорова, Н. А. Анализ финансовой отчетности: Учебник. — 7-е изд. — М.: ДИС, 2009.
3. Уфимцев, М. В. Методы анализа данных. — М.: МАКС Пресс, 2007.
4. Шлее, М. Профессиональное программирование на C++.QT4. — СПб: БХВ, 2006.
5. Никифорова, Н. А., Донцов, Е. В. Оперативный мониторинг финансового состояния // Управленческий учет. — 2009. — № 6.
6. Воронцов, К. В. Лекции по методу опорных векторов [Электронный ресурс] / Сайт Вычислительного центра им. А. А. Дородницына РАН. — Режим доступа: http://www.ccas.ru/voron/download/SVM.pdf.
7. Чубукова, И. А. Учебный курс Data Mining [Электронный ресурс] / Интернет-Университет Информационных Технологий (ИНТУИТ). — Режим доступа: http://www.intuit.ru/department/database/datamining.
8. Muller, K., Mika, S. An Introduction to Kernel-Based Learning Algorithms // IEEE Neural Networks. — 2001. — № 12 (2). — P. 181–201.
9. Platt, J. C. Fast Training of Support Vector Machines using Sequential Minimal Optimization // Advances in Kernel Methods — Support Vector Learning / B. Schölkopf, C. Burges, and A. Smola, eds., — MIT Press, 1999. — P. 41–65.
10. Vapnik, V. N. An Overview of Statistical Learning Theory // IEEE transactions on neural networks, vol. 10. — 1999. — № 5. — P. 988–999.
11. Мерков, А. Б. О статистическом обучении [Электронный ресурс] / Лаборатория распознавания Московского Центра непрерывного математического образования. — Режим доступа: http://www.recognition.mccme.ru/pub/RecognitionLab.html/slt.html.
12. Platt, J. C. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines [Электронный ресурс] / Microsoft Research. — Режим доступа: http://research.microsoft.com/en-us/um/people/jplatt/smotr.pdf.
13. Лукьяница, А. А., Шишкин, А. Г. Цифровая обработка видеоизображений. — М.: Ай-Эс-Эс Пресс, 2009.
14. Advances in Knowledge Discovery and Data Mining / Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusam, eds. — MIT Press, 1996.