Мукашев И.Е.

Применение диффузных карт для оценки кредитоспособности

(СПбГУЭФ, Санкт-Петербург)

Существенной особенностью процесса оценки платежеспособности заемщика и уровней риска в банковских информационных системах является наличие неопределенностей на различных стадиях технологии оценки уровня риска информационными системами в банках.

Технологии оценки, используемые в настоящее время в банковской практике, часто основаны на формальном аппарате, который не учитывает особенностей оценки платежеспособности и предназначен для условий, достаточно далеких от тех, в которых решаются задачи кредитования, что существенно снижает их эффективность.

База знаний по кредитозаемщикам позволяет применить технологию классификации с обучением. Задача обучения сводится к разбиению пространства признаков на классы. Процедура самообучения (обучение без эксперта) основана на автоматической классификации.

Технологии классификации и выборки представлены основным набором алгоритмов: AdaBoost, SVM, Нейронные сети, Linear Discriminate Analysis.

На основании обзора литературных источников [2-4], можно определить два подхода к классификации на основе обучения с экспертом. Первый из них – двухэтапный алгоритм разделения данных на группы, первый этап из которого известен как K-means, а второй – Support Vector Machines (SVM – Support Vector Mashine [5], [6]). Второй метод основан на сингулярном разложении матрицы (SVD-Singular value decomposition [6]).

Таким образом, решается задача классификации, где исходя из имеющейся в базе данных информации, обработанной на стадии обучения, получается функция, наиболее точно разделяющая выборку клиентов на «плохих» и «хороших».

Очевидны проблемы, возникающие при имеющихся технологиях классификации заемщиков: нечеткость определения классов, отсутствие строгого разделения и пересечение классов, подобие сингулярных векторов, представляющих классы, отсутствие непрерывной меры платежеспособности, зависимость результатов от обучающей базы.

Опишем модель представления многомерной информации. Данные из базы структурируются в виде вектора представлений с равноценными координатами, измеряемые в одной шкале.

Априори известно, что все объекты, содержащиеся в базе данных, поделены на две категории. Применительно к нашему случаю, заемщики поделены на «плохих» и «хороших». Наша задача – выяснить закономерности в представлении, которые разделяют две названные группы.

При решении этой проблемы мы будем опираться на метод «диффузных карт», описанный в [4].

Этот метод впервые применялся для моделирования трехмерных объектов на базе множества представлений объекта двумерными проекциями (фотографиями). Суть метода заключается в том, что многомерные данные проецируются в математическое многообразие малой размерности с сохранением взаимных отношений между данными. При этом топология многообразия моделирует различие между проекциями. То есть, вариация данных описывается многообразием, выстраиваемым диффузной картой. В случае, когда многообразие трехмерно, оно является трехмерной моделью проекций.

В нашем случае мы должны выделить такие показатели, при которых выстраивается диффузная карта, в которой названные два кластера («положительные клиенты» и «отрицательные клиенты») разделятся явно. Отбор показателей осуществляется перебором. Для каждого набора мы опишем диффузный процесс, который выявит скрытые закономерности, существующие между характеристиками, разделяющие две группы.

Модель, которую мы предлагаем, основывается на случайном блуждании по графу.

Рассмотрим I – множество объектов, представленных многомерными векторами. Построим граф G=(V, E), где множество вершин V соответствует объектам, принадлежащим I, а множество ребер – мера локальной близости между векторами, индуцированной нормой L_2.

Далее мы инициализируем случайное блуждание точки по графу G. Блуждание по графу происходит в областях сгущения плотности, поскольку вероятности перехода из узла в узел в плотных участках больше, чем переход из одной точки сгущения в другую. Поэтому такое блуждание выделяет кластеры, как области наиболее вероятного нахождения точки при случайном блуждании.

Таким образом, случайное блуждание разделяет всю область V на отдельные кластеры, которые обусловлены скрытыми взаимосвязями между элементами множества I.

Теперь мы дадим более формальное аксиоматическое определение весовых функций, связанных с ребрами графа G.

Весовая функция ω_ε определяется как такое отношение локальной близости между вершинами графа, которое обладает следующими свойствами: симметричностью, неотрицательностью, свойством разреженности. Как правило, в качестве функции ω_ε выбирается «гауссовское ядро» [6].

Теперь опишем формально случайный процесс блужданий по графу G.

Определим вес каждой вершины в графе d(x), как сумму значений функции ω_ε по всем парам (x, y).

Нормализуем весовые функции ω_ε как строки стохастической Марковской матрицы [1]. Более формально, рассмотрим матрицу P, составленную из p(x, y), где p(x, y) – отношение ω_ε (x, y) к d(x).

Величина p(x, y) может быть интерпретирована, как вероятность перехода из точки x в точку y за один шаг. Определим теперь вероятность перехода из x в y за время t как $p_t(x, y)$. Матрица, составленная из элементов $p_t(x, y)$, задает разбиение графа на кластеры при стремлении параметра t к бесконечности.

Как было показано в [7], диффузное расстояние $P_t(x,y)$ в квадрате может быть вычислено как сумма произведений $\lambda_j^{2t}(\psi_j(x) - \psi_j(y))^2$, где j=1, …, m, $\lambda_1, \lambda_2, .., \lambda_m$ – собственные числа матрицы P, а $\psi_1, \psi_2,\ldots \psi_m$ – соответствующие им собственные векторы.

Отбросив теперь члены при собственных числах, близких к 0 и оставив первые w самых существенных слагаемых, приходим к тождественному равенству между $P_t(x,y)$ и суммой произведений $\lambda_k^t(\psi_k(x) - \psi_k(y))$.

Определим отображение

$$\psi_k(x)= (\lambda_1^t \psi_1(x), \ldots, \lambda_w^t \psi_w(x)).$$

Нетрудно видеть, что оно обладает следующими свойствами:

- отображение происходит в пространство размерности w;
- отображение не является линейным;
- расстояние между образами точек равно диффузному расстоянию, то есть, вероятности попасть из точки x в точку y при случайном блуждании по графу G за время t.

Данное отображение будем называть «диффузной картой».

Рисунки 1 и 2 иллюстрируют разделение кластеров после применения диффузных карт.

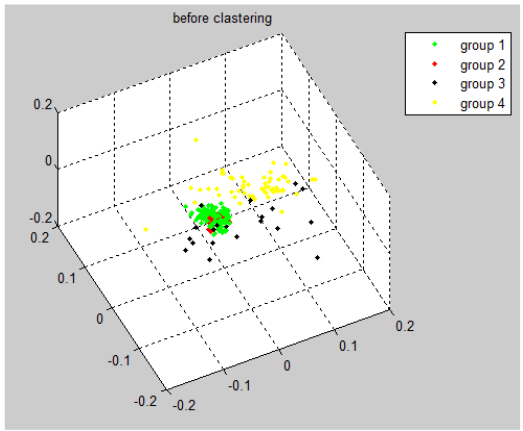

Рис. 1. Визуализация трех случайно выбранных показателей

Видно, что в результате разнонаправленности воздействия показателей кластеры полностью перемешаны и классификация по ним невозможна.

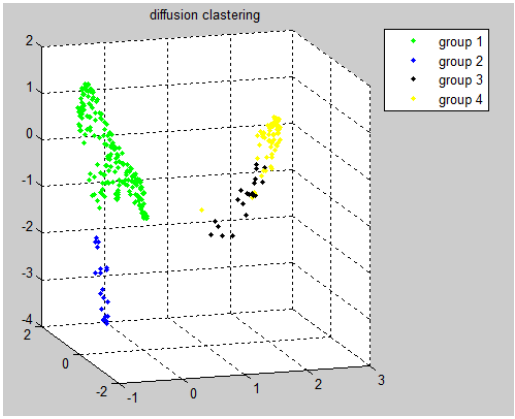

Рис. 2. Визуализация показателей после применения диффузных карт

Техника предусматривает отображение показателей в виртуальное пространство отношений, и после понижения размерности, пространство показателей проецируется в трехмерное пространство, где образы показателей, относящихся к одному классу, хорошо разделены.

По рисунку 2 можно отметить, что заемщики по заданным признакам в результате применения диффузных карт поделились на 4 категории, которые могут иметь логическое осмысление, но это выходит за рамки данной статьи.

Таким образом, данные о клиентах из обучающей базы при помощи представления показателей многомерным вектором, а затем отображение этого вектора в маломерное пространство, представляются набором точек в этом пространстве. Методика такого отображения предполагает явное линейное разделение представлений клиентов между «плохими» и «хорошими».

Для разделения кластеров и выявления взаимосвязей между данными, а также представления кластеров в пространстве малой размерности была применена техника случайных процессов в Марковских цепях.

Литература

1. Норвич А.М., Туркcен И.Б. Построение функций принадлежности // Нечеткие множества и теория возможностей. – М.: Радио и связь, 1988. – С. 64-71.

2. Averbuch Amir Z., Zheludev Michael V. Target recognition in hyperspectral images, School of Computer Science Tel Aviv University, Tel Aviv 69978, 2010, Israel.

3. Bayliss Jessica D., Gualtieri J. Anthony and Cromp Robert F. Analyzing hyperspectral data with independent component analysis. In Proc. of the SPIE conference 26th AIPR Workshop: Exploiting New Image Sources and Sensors, volume 3240, pages 133-143, 1997.

4. Coifman Ronald R., Lafon Stephane M. Diffusion maps, Appl. Comput. Harmon. Analysis, Mathematics Department, Yale University, New Haven, CT 06520, USA.

5. Hyvarinen Aapo, Karhunen J. and Oja E. Independent Component Analysis. John Wiley& Sons, Inc., 2001.

6. Raged R., Cupta M. Fuzzy set theory introduction.//In: Fuzzy Automata and Decission Processes/ Ed. by Gupta M., Saridis G., Gaines B. – Amsterdam: Nord-Holland, 1977, p.105—131.

7. Zheludev Michael V. Classification with diffusion maps. Computer Science Department, Technion – Israel Institute of Technology, VULCAN 28.04.10.

Сотавов А.К.

Формализация правил извлечения знаний о рыночном позиционировании результатов инновационной деятельности

(СПбГУЭФ, Санкт-Петербург)

В современных условиях экономический рост отождествляется, прежде всего, с научно-техническим прогрессом, интеллектуализацией основных элементов производства. Новые знания находят воплощение в технологиях, оборудовании, управлении и организации производства. На их долю в развитых странах приходится 70-80% от прироста ВВП. Кроме того, значительная часть ВВП формируется через каналы внешней торговли с помощью иностранных инвестиций и технологий. Однако процессы, происходящие в научно-промышленной сфере России, находятся в серьезном противоречии с практикой мирового сообщества, в которое наша страна стремится интегрироваться посредством рыночных реформ и открытой экономики. Эта проблема усугубляется с учетом того факта, что взаимодействие с внешним миром стало столь интенсивным, что уже не может рассматриваться в качестве фактора приспособления отечественной экономики. Важной составляющей российской экономики становятся связи с региональными группами, другими странами, а также с международными экономическими и финансовыми организациями.

Привлечение иностранных компаний к российской инновационной деятельности уже стало традиционным, например, участие в отдельных стадиях инновационных процессов таких зарубежных фондов, как: фонды Сороса и Форда, Европейского инструмента соседства и партнерства, фонда гражданских исследований и разработок США, Американского фонда инженерной информации и многих других. В связи с кризисом банки в системах ипотечного и потребительского кредитования стали проявлять больший интерес к инвестированию инновационных проектов, некоторые из