Применение байесовского классификатора для оценки надежности банка

© 2010 С.И. Пшеничный
Финансовая академия при Правительстве Российской Федерации
E-mail: pshen2010@yandex.ru

Байесовские сети хорошо зарекомендовали себя в задачах на проведение классификации данных и наблюдений. Байесовский классификатор можно использовать для целей определения надежности банка, что может позволить в большей мере владеть информаций при выборе банка. В данной работе дано определение надежности банка с точки зрения вероятностного подхода, представлен вид наивного байесовского классификатора на основе байесовской сети для определения надежности банка и приведены заключения о применимости его при нарушении исходных допущений. Перечислены методы повышения точности классификации банков.

Ключевые слова: надежность банка, байесовские сети, наивный байесовский классификатор, допущения наивного байесовского классификатора.

В байесовском подходе предполагается, что случайность есть мера нашего незнания происходящих явлений. Таким образом можно интерпретировать любое случайное событие.

Идея байесовского подхода заключена в переходе от априорных знаний о системе к апостериорным знаниям с учетом принятия во внимание наблюдаемых участвующих в расчетах показателей.

В байесовском подходе все величины и параметры считаются случайными, так как, даже если есть их конкретные значения, точные законы распределения неизвестны. Оценить неизвестный параметр системы означает найти апостериорное его распределение.

Случайность в выборе надежного или ненадежного банка характеризует незнание полного информационного поля системы. Главные пробелы в знаниях связаны со сложностью такой системы, как банк. Существуют пробелы в знаниях о механизмах взаимовлияния факторов, об их совместном или противоположном воздействии на состояние надежности банка.

Надежность банка - это вероятность банка быть надежным, т.е. вероятность того, что банк выполнит взятые на себя обязательства. В рамках данного определения надежностью банка полагается его нахождение в конкретном состоянии - состоянии “надежного банка”.

Так, надежность коммерческого банка будет определенным качеством банка. Можно проследить проявление данного качества в зависимости от значения других показателей и качеств банка. Иными словами, надежность банка будет исследоваться на определенных условиях.

Ключевым понятием в методе расчета вероятности надежности коммерческого банка с помощью байесовского подхода выступает байесовская сеть. Байесовская сеть, или байесовская сеть доверия, - это вероятностная модель, представляющая собой множество переменных и их вероятностных зависимостей. Байесовские сети доверия визуализируются как направленный ациклический граф. Байесовская сеть рассматривает все множество факторов системы в их взаимосвязях. На плоскости все факторы можно изобразить как узлы-вершины, а связи рассматриваемой системы могут быть отражены ребрами, соединяющими вершины в направлении влияния.

Качество работы сети зависит от топологии сети, количества узлов, от правильно подобранных параметров распределений случайных факторов.

В контексте моделирования надежности коммерческого банка показатель надежности выступает в качестве целевого элемента системы. Пример байесовской сети для определения надежности банка представлен на рисунке.

В построенной байесовской сети наивный байесовский классификатор принимает следующий вид:

$$p(\text{надежен} \mid \text{параметр}_1, \ldots, \text{параметр}_n) =$$

$$= \frac{1}{P(\text{параметр}_1 \ldots \text{параметр}_n)}$$

$$p(\text{надежен}) \prod_{i=1}^{n} p(\text{параметр}_i \mid \text{надежен}),$$

где параметр_i - реализация каждой из n выбранных переменных.

В наивном байесовском классификаторе делается строгое предположение о независимости факторов между собой. Преимуществом наивно-

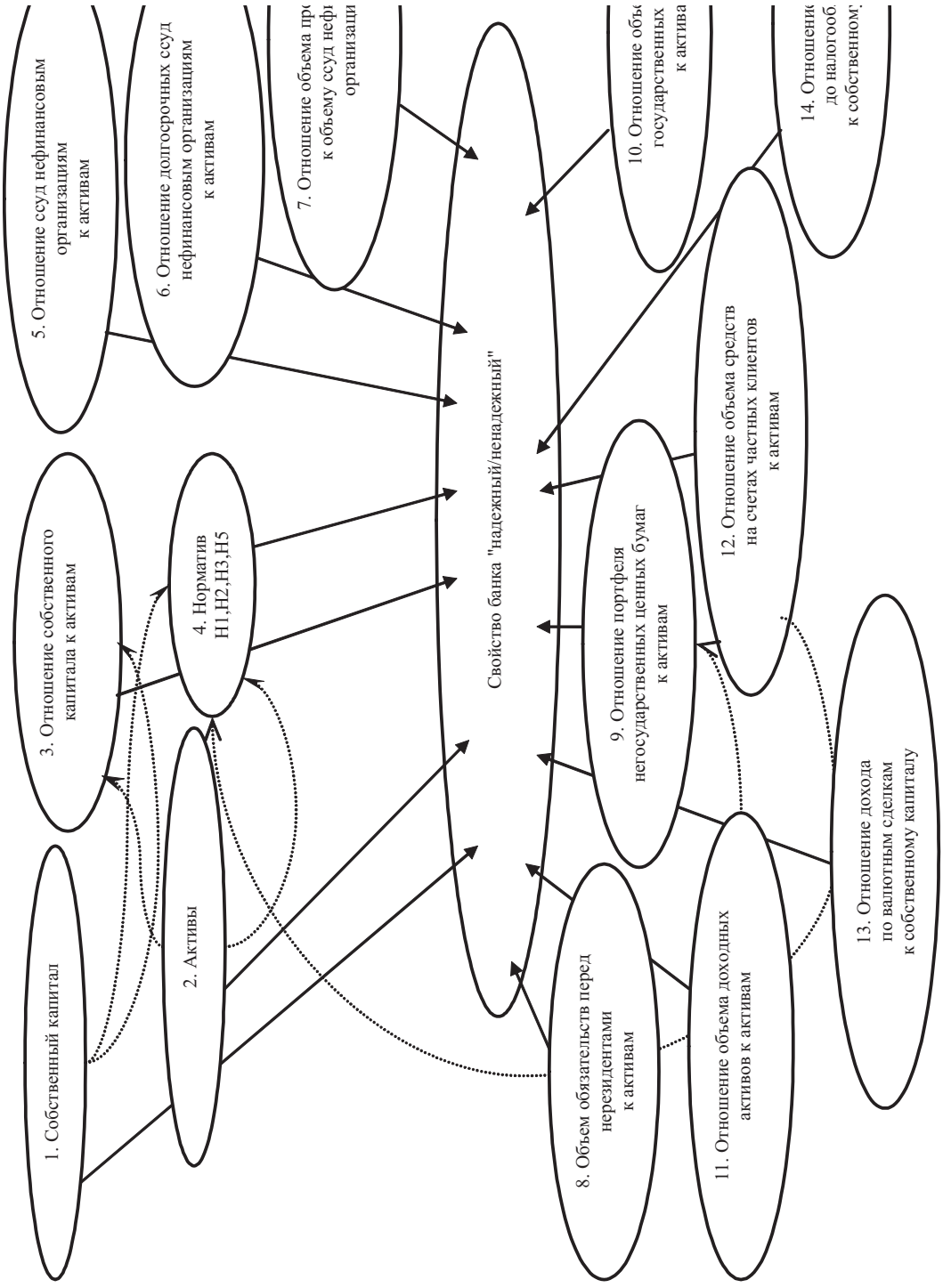

Рис. **Топология байесовской сети для надежности банка**

го байесовского классификатора является требование к размеру выборки. Существенным моментом в данном классификаторе выступает то, что его можно применять, когда выполняется предположение о независимости между факторами. В противном случае результат применения классификатора может оказаться неточным, и сумма вероятностей принадлежности к обоим классам не будет равняться 1. Чтобы определить эти вероятности принадлежности к классам, необходимо будет проводить нормировку полученных результатов.

Наивный байесовский классификатор является одним из самых эффективных способов обработки данных. Результаты классификации, полученные при применении наивного байесовского классификатора, могут удивлять, так как в повседневной жизни редко встретится ситуация, удовлетворяющая предположению о независимости параметров модели. Тем не менее, наивный байесовский классификатор достаточно широко используется во многих областях[1].

Логичным следствием таких практических применений будет вопрос, почему наивный байесовский классификатор применим во многих случаях, даже когда наблюдаются существенные зависимости между переменными.

Обозначим как вектор $E = (F_1 \quad \ldots \quad F_n)$ набор значений всех рассматриваемых факторов.

Согласно теореме Байеса вероятность быть надежным банком при имеющихся значениях факторов будет равна:

$$p(C = надежен | E) =$$

$$= \frac{p(C = надежен)\, p(E | C = надежен)}{p(E)}.$$

Конкретное наблюдение E признается “надежным”, если байесовский классификатор

$$f_b(E) = \frac{p(C = надежен | E)}{p(C = ненадежен | E)} \geq 1.$$

Учитывая условия независимости факторов:

$$p(E | C = надежен) = p(F_1 \ldots F_n | C = надежен) =$$

$$= \prod_{i=1}^{n} p(F_i | C = надежен).$$

Получаем наивный байесовский классификатор

$$f_{nb}(E) =$$

$$= \frac{p(C = надежен | E)}{p(C = ненадежен | E)} \prod_{i=1}^{n} \frac{p(F_i | C = надежен)}{p(F_i | C = ненадежен)} \geq 1.$$

Совместная вероятность для сети на рисунке с учетом зависимостей:

$$p(F_1, \ldots, F_n, C = надежен) =$$

$$= p(C = надежен) \prod_{i=1}^{n} p(F_i | par(F_i), C = надежен),$$

где $par(F_i)$ - фактор родительского узла для i-го фактора.

Для каждой пары факторов, где проявляется зависимость, можно для самой переменной и ее родителя записать локальные выражения зависимости по выбранным классам.

$$dp^{надежный}(F_i | par(F_i)) = \frac{p(F_i | par(F_i), C = надежный)}{p(F_i | C = надежный)}.$$

$$dp^{ненадежный}(F_i | par(F_i)) = \frac{p(F_i | par(F_i), C = ненадежный)}{p(F_i | C = ненадежный)}.$$

Величина $dp^{надежный}(F_i | par(F_i))$ характеризует силу зависимости между двумя факторами в классе надежных баков.

Величина $dp^{ненадежный}(F_i | par(F_i))$ характеризует силу зависимости между двумя факторами в классе ненадежных баков.

Для любого узла F_i из расширенной байесовской сети, т.е. сети, в которой наблюдаются зависимости между переменными, можно составить коэффициент локальных выражений зависимости:

$$k(F_i) = \frac{dp^{надежный}(F_i | par(F_i))}{dp^{ненадежный}(F_i | par(F_i))}.$$

Коэффициент дает численное объяснение степени влияния локальной зависимости в байесовской сети на итоговую классификацию.

Если F_i является независимым фактором, то $k(F_i)$ равен 1.

Если $dp^{надежный}(F_i | par(F_i)) = dp^{ненадежный}(F_i | par(F_i))$, то $k(F_i)$ равен 1. Это означает, что распределение влияния зависимости на классификацию распределяется равномерно на оба класса. Зависимость не будет влиять на классификацию, неважно, насколько существенной она является.

Если $k(F_i) \succ 1$, то зависимость между факторами будет оказывать больший эффект на классификацию банка как надежного, чем как ненадежного, и наоборот, если $k(F_i) \prec 1$.

Коэффициент $K(E)$ определяет различие между байесовским классификатором и наивным байесовским классификатором. Если $K(E)$ равен 1, то результаты классификации совпадают.

Но равенство 1 необязательно. Согласно проведенному анализу [2] при известных $E = (F_1 \quad ... \quad F_n)$ байесовский классификатор $f_b(E)$ и наивный байесовский классификатор $f_{nb}(E)$ для построенной сети будут эквивалентны тогда и только тогда, когда $f_b(E) \geq 1$, $K(E) \leq f_b(E)$ или когда $f_b(E) \prec 1$, $K(E) \succ f_b(E)$. Разный результат классификации может быть только при несоблюдении указанных условий. Это условие является необходимым и достаточным условием совпадения байесовского классификатора и наивного байесовского классификатора, а также оптимальности второго при применении в системах с проявляющейся зависимостью факторов между собой.

Качество классификатора определяется количеством результатов применения классификатора, совпадающих с действительностью, т.е. чем больше правильных классификаций было получено, тем выше качество классификатора[3]. Таким образом, при соблюдении описанных ранее условий качества байесовского классификатора и наивного байесовского классификатора близки.

Существуют также методы предобработки данных перед применением байесовского классификатора, позволяющие получить более точное распределение по классам.

К таковым относятся методы фильтрации, методы свертывания и дискретизация первоначальных данных.

В методе фильтрации каждому фактору приписывается уровень значимости.

$$p(C = \textit{надежен} \mid E) =$$

$$= \frac{p(C = \textit{надежен}) \prod_{i=1}^{n} p(F_i \mid C = \textit{надежен})^{\omega_i}}{p(E)},$$

где ω_i - весовой коэффициент, характеризующий значимость рассматриваемого фактора при проведении процедуры классификации.

Сам метод сводится к оцениванию данных весов. Для факторов, которые подозреваются в нарушении предположения о независимости, устанавливаются меньшие веса, чем для факторов независимых вкладов в классификацию.

Метод свертывания может из двух зависимых друг от друга факторов выбирать один или заменять оба фактора на общий составной показатель. В силу самой специфики построения метода свертывания он приводит к улучшению наивного байесовского классификатора. Процедура обработки входящей информации может производиться до тех пор, пока последующее изменение уже не будет давать улучшения точности классификатора.

На первом этапе метод выделяет факторы с наибольшей мерой зависимости. Строится коэффициент детерминации R^2.

Каждое проявление зависимости характеризуется следующим элементом:

$$d_i = (A_i, A_j, R(A_i, A_j)).$$

Составляется список всех пар зависимостей в порядке убывания меры зависимости. Далее удаляются из списка факторы с наибольшей мерой зависимости.

Вторым этапом для каждого фактора, не удаленного на первом этапе, строится свой список с соответствующей оценкой влияния фактора на точность метода классификации. Для каждого фактора оценивается изменение точности разбиения на классы при удалении его из списка. Если при удалении из списка рассматриваемого фактора точность классификатора увеличивается или остается той же, то итоговый список не будет содержать данный фактор.

Наибольшее распространение среди способов улучшения байесовского классификатора получили алгоритмы дискретизации входных данных. Вместо вычисления вероятностей, используя нормальный закон распределения, различными алгоритмами можно провести дискретизацию данных. К таким относятся разбиение непрерывного показателя на *k* равных интервалов или разбиение на интервалы таким образом, чтобы в каждом интервале было одинаковое количество наблюдений.

Проведенные исследования по сравнению подобных алгоритмов[4] показали, что байесовский классификатор с применением дискретизации параметров дает лучшие результаты, чем расчеты, основанные на предположении о нормальном законе распределения непрерывных факторов. Среди многих таких алгоритмов выделяется алгоритм, основанный на снижении энтропии.

Таким образом, используя априорную информацию о состоянии банков и данные офи-

циальных отчетностей, можно выбрать ряд показателей, которые будут оказывать существенное влияние на надежность банка. Используя методы фильтрации, свертывания и дискретизации, можно построить байесовскую сеть, которая будет позволять проводить классификацию банков на надежные и ненадежные с высоким качеством.

[1] *Zhang H.* The Optimality of Naive Bayes. Faculty of Computer Science University of New Brunswick Fredericton, New Brunswick, Canada.

[2] *Kononenko I.* Comparison of inductive and naive Bayesian learning approaches to automatic knowledge acquisition // Wielinga B. ed. Current Trends in Knowledge Acquisition. IOS Press, 1990.

[3] *Pazzani M.J.* Search for dependencies in Bayesian classifiers // In Fisher D., Lenz H.J. eds. Learning from Data: Artificial Intelligence and Statistics V. Springer Verlag, 1996.

[4] *Kotsiantis S.B., Pintelas P.E.* Increasing the Classification Accuracy of Simple Bayesian Classifier. Educational Software Development Laboratory. Department of Mathematics. University of Patras, Hellas.

Поступила в редакцию 09.01.2010 г.