Сегментация персональных данных с использованием показателя сопряженности

Д.А. Жердев[1], П.В. Хрипунов[2]

[1]Самарский национальный исследовательский университет им. академика С.П. Королева, Московское шоссе 34А, Самара, Россия, 443086
[2]Пенсионный фонд Российской Федерации, Шаболовка 4, Москва, Россия, 119991

Аннотация. В работе рассматривается возможность применения метода обработки персональных данных и выявлений сходства записей на основе разделения большой базы данных на подклассы. Для решения поставленной задачи нами используется показатель сопряженности. Данный показатель демонстрирует свою эффективность как при решении задачи распознавания, так и при решении задачи кластеризации данных.

1. Введение

Проблема понимания и анализа большого объема данных (data mining) является весьма актуальной ввиду обработки огромного количества информации с использованием электронно-вычислительных машин. В данной исследовательской области находят свое применение различные методы, которые также широко используются в смежных областях, например, распознавание образов, тематической классификации, кластеризация изображений и т.д. Существует множество работ по теме кластеризации больших данных [1]-[4]. В данной работе рассматривается возможность классификации входного вектора данных представляющего из себя заданную строку, содержащую персональные данные потенциального клиента. Классификатор должен с некоторой вероятностью определить имеется ли схожий элемент в общей базе данных клиентов, которая была предварительно сегментирована на отдельные классы. Для достижения данной целимы используем показатель сопряженности. Эффективность использования данного показателя была показана в задачах распознавания лиц на изображениях [5], а также объектов на радиолокационных изображениях [6],[7].

2. Сегментация и классификация персональных данных

В данном разделе описывается подход кластеризации персональных данных на основе использования показателя сопряженности. В зависимости от величины показателя сопряженности текущего вектора с матрицей класса можно судить о принадлежности вектора к определенному классу. Чем выше значение показателя, тем больше вероятность, что текущий вектор схожс остальными векторами класса, которые формируют матрицу класса:

$$\mathbf{X}_k = \left[\mathbf{x}_1(k), \mathbf{x}_2(k), ..., \mathbf{x}_j(k), ..., \mathbf{x}_M(k) \right], k = \overline{1,K},$$

где $\mathbf{x}_j = \left[x_1, x_2, ..., x_i, ..., x_N \right]$ – это $N\times1$-вектор признаков.

Сам показатель сопряженности записывается в виде:

$$R_k\left(\mathbf{x}_j\right) = \frac{\mathbf{x}_j^T \mathbf{Q}_k \mathbf{x}_j}{\mathbf{x}_j^T \mathbf{x}_j}, \quad k = \overline{1,K},$$

где K – число распознаваемых классов,

$$\mathbf{Q}_k = \mathbf{X}_k \left[\mathbf{X}_k^T \mathbf{X}_k \right]^{-1} \mathbf{X}_k^T, \quad k = \overline{1,K},$$

– $N \times N$ матрица k -го класса.

Для формирования числового вектора из одной выбранной записи из базы данных, которая представляет себя значения категориального типа, используется кодирование буквы индексом. Каждая буква русского алфавита А-Я кодируется числами 1-33. В результате формируется «новая» база данных векторов в числовом представлении.

В базе данных три поля: фамилия, имя, отчество. Для удобства максимальное количество возможных символов в базе данных было выбрано равным 100 и в результате все вектора «новой» базы данных получились размером в 300 компонент. Все вектора в базе данных должны быть одинакового размера. Если размер одного поля, например, 12 символов, то первые 12 компонент вектора заполняютсячисленным представлением, характеризующим поле, а остальные 88 компонент равны нулю.При работе с конкретной базой данных все персональные данные были зашифрованы, т.е. сложены с некоторым цифровым ключом.

Для разбиения классов на подклассы мы используем схожую процедуру, которая применялась в работе [6] для кластеризации радиолокационных изображений. На первом шаге из всего множества, например M, векторов выбираются два наиболее «удалённых» вектора, которые можно обозначить как $\mathbf{x}_1$, $\mathbf{x}_M$. Для данных векторов величина нормированного коэффициента корреляции должна быть минимальна.

Далее в алгоритмеиз оставшегося множества векторов присоединяем к двум найденным на первом шаге векторам по одному вектору ($\mathbf{x}_2$, $\mathbf{x}_{M-1}$), для которых величины коэффициента корреляции

$$R_{1,2} = \frac{\left\langle \mathbf{x}_1^T \mathbf{x}_2 \right\rangle}{\left\|\mathbf{x}_1\right\|\left\|\mathbf{x}_2\right\|},$$

$$R_{M-1,M} = \frac{\left\langle \mathbf{x}_{M-1}^T \mathbf{x}_M \right\rangle}{\left\|\mathbf{x}_{M-1}\right\|\left\|\mathbf{x}_M\right\|},$$

принимают максимальные значения. В результате полученные пары векторов $\mathbf{x}_1$, $\mathbf{x}_2$ и $\mathbf{x}_{M-1}$, $\mathbf{x}_M$ образуют два подпространства, которые формируют матрицы $\mathbf{X}_{1,2}$ и $\mathbf{X}_{M-1,M}$ соответственно. Далее, согласно алгоритму, из оставшегося множества векторов к уже образованным подпространствам присоединяются два вектора $\mathbf{x}_3$, $\mathbf{x}_{M-2}$ ближайшие к этим двум подпространствам по критерию максимумапоказателя сопряжённости.

Так как база данных содержит большое количество векторов процесс продолжается до нахождения определенного числа векторов в обоих подпространствах. На примере исходной базы данных конечное искомое подпространство составило 15 векторов в каждой матрице $\mathbf{X}_k, \mathbf{X}_l$,которые соответствуют двум подклассам.

Процедура, описанная выше, со всеми «неразмеченными» векторами повторяется циклически. Разбиение на подпространства продолжается до тех пор, пока все вектора не окажутся в любом из подпространств. На этапе распознавания при определенном решающем правиле, векторближайший к одному из образованных описанным способом подклассов, считается принадлежащим исходному классу.

3. Результаты и обсуждение

В данной работе ставиться задача возможности определения существует ли некоторая заданная запись в базе данных. После процесса кластеризации, разбиения большой базы данных на некоторые подклассы можно узнать принадлежность входного вектора к некоторому классу. Подкласс хранит малое количество векторов по сравнению со всей базой, таким образом, после классификации текущего вектора легко будет произвести анализ данных в подклассе и выявить насколько доступно вносить в базу новое значение.

Таким образом, для проверки поставленного выше предположениябыл поставлен следующий эксперимент. Из базы данных в 1041100 записей были произведены пятьслучайных выборки в 1040 записей. Каждая такая выборка была поделена на 80 подклассов, в каждом

подклассе содержится по 13 векторов. После проведенной процедуры кластеризации были выполнена классификация генерируемых векторов.

Входные векторы классификатора формировались на основе уже имеющихся в конкретной выборке данных. Например, для записи, представленной на рисунке 1, добавлялись ошибочные вхождения, некоторые буквы заменены,удалены или добавлены лишние. Такое может происходить, например, при автоматизированной обработке отсканированных рукописных документов, либо при некорректном заполнений электронных документов.

Рисунок 1.Пример преобразования данных в некорректные.

В эксперименте по классификации были проверены 20 векторов описанного выше вида. Все вектора были безошибочно классифицированына основании подхода один ко многим [8] расширяющим возможности бинарной классификации алгоритма опорных подпространств [7]. Кроме того, среднее значение величины показателя сопряженности для верно определенного класса составляет 0,95. Данный факт несомненно говорит о надежности использовании показателя сопряженности в задачах такого рода. Это является позитивным моментом для продолжения исследований такого рода как на базах данных более сложного вида, с большим числом полей, так и для классификации на полной базе данных из миллиона и более записей.

4. Литература

[1] Yang, Y. CLOPE: a fast and effective clustering algorithm for transactional data / Y. Yang, Guan, J.You // Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM, 2002. – P. 682-687.

[2] Zhang T. BIRCH: an efficient data clustering method for very large databases / T. Zhang, R. Ramakrishnan, M. Livny //ACM Sigmod Record. – 1996. – Vol. 25(2). – P. 103-114.

[3] He, Z. A cluster ensemble method for clustering categorical data / Z. He, X. Xu, S. Deng // Information Fusion. – 2005. – Vol. 6(2). – P. 143-151.

[4] Huang, Z. A fast clustering algorithm to cluster very large categorical data sets in data mining / Z. Huang // DMKD. – 1997. – Vol. 3(8). – P. 34-39.

[5] Козин, Н.Е. Построение классификаторов для распознавания лиц на основе показателей сопряженности / Н.Е. Козин, В.А. Фурсов // Компьютерная оптика. – 2005. – Т. 28. – С. 160-163.

[6] Фурсов, В.А. Построение опорных подпространств в задачах распознавания фрактальных изображений / В.А.Фурсов, Е.Ю. Минаев // Труды международной научной конференции «Информационные технологии и нанотехнологии (ИТНТ-2016)», 2016. – С. 530-537.

[7] Жердев, Д.А. Распознавание объектов на радиолокационных изображениях с использованием показателей сопряжённости и опорных подпространств / Д.А. Жердев, Н.Л. Казанский, В.А. Фурсов // Компьютерная оптика. – 2015. – Т. 39, № 2. – С. 255-264.

[8] Bishop, Ch.M. Pattern Recognition and Machine Learning / Ch.M. Bishop. – New York: Springer, 2006. – 738 p.

Personal data segmentation based on conjugation index usage

D.A. Zherdev[1], P.V. Hripunov[2]

[1]Samara National Research University, Moskovskoye shosse, 34, Samara, Russia, 443086
[2]Pension Fund of the Russian Federation, Shabolovka 4, Moscow, Russia, 119991

Abstract. The paper proposes a method for processing personal data that allows them to be divided into many segments or classes. The customer database is used as the source data. We use the indicator of conjugacy that has already proved the effectiveness in both recognition and clustering of data problems.

Keywords: clustering, data analyse, data mining.