УДК 615.471:616-071

И.Г. Перова, Е.В. Бодянский

Харьковский национальный университет радиоэлектроники, Харьков

НЕЧЕТКАЯ КЛАССИФИКАЦИЯ ДАННЫХ МЕДИКО-БИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ В УСЛОВИЯХ ДЕФИЦИТА ИНФОРМАЦИИ

В статье рассмотрен подход, позволяющий подвергать процедуре нечеткой кластеризации-классификации выборки медицинских данных, сильно ограниченные по объему с помощью метода нечеткой пространственной экстраполяции. Предложенная процедура относится к направлению Medical Data Mining и представляет собой гибридную систему, позволяющую решать задачу диагностирования разного рода заболеваний в условиях ограниченной выборки, полного или частичного перекрытия классов, разной их плотности, разного численного наполнения и требующей для своего обучения малых объемов априорной информации.

Ключевые слова: *нечеткая кластеризация-классификация, дефицит информации, нечеткая пространственная экстраполяция.*

Введение

В настоящее время методы интеллектуального анализа данных (Data Mining) [1] получили широкое распространение для решения широкого класса задач в промышленности, финансах, банковской сфере, сельском хозяйстве. Не обошли стороной эти методы и медицинские приложения, где в силу специфики решаемых задач сформировалось направление, известное как Medical Data Mining [2]. Основной задачей, решаемой в рамках Medical Data Mining, являются задачи диагностирования, решаемые на основе парадигм как обучения, так и самообучения и с математической точки зрения сводящиеся либо к решению задач классификации (распознавания образов), либо кластеризации. Особенностью этого направления является невозможность использования традиционных методов Data Mining в чистом виде, что связано с целым рядом обстоятельств: прежде всего это ограниченность выборок, подлежащих классификации, существенное перекрытие классов, относящихся к разным заболеваниям, нелинейный характер разделяющих гиперповерхностей, наличие аномальных наблюдений – выбросов, искажающих исходную информацию, значительная роль субъективного человеческого фактора, не обеспечивающая получение точных данных. Все вышеперечисленные обстоятельства приводят к формированию невыпуклых и нечетких классов, при этом адекватным математическим аппаратом для работы с такими данными являются методы вычислительного интеллекта и, прежде всего, искусственные нейронные сети [3 – 5], системы нечеткого вывода [6] и гибридные системы, построенные на их основе – нейро-фаззи-системы [7, 8]. Здесь следует отметить, что все равно в этом классе задач данные системы не являются панацеей, поскольку для своего обучения требуют больших объемов информации, которыми зачастую врачи не располагают.

Метод нечеткой пространственной экстраполяции для задач кластеризации-классификации

Аппаратом, решающим задачи классификации в условиях малых объемов выборок, является метод линейной пространственной экстраполяции [9], однако здесь надо отметить, что сам по себе этот метод является линейным, то есть разделяющая гиперплоскость является линейной, а также он является совершенно четким. В связи с этим представляется целесообразным на основе этого метода, а также методов вычислительного интеллекта провести синтез гибридной системы вычислительного интеллекта, позволяющей решать задачу диагностирования разного рода заболеваний в условиях ограниченной выборки, полного или частичного перекрытия классов, разной их плотности, разного численного наполнения и требующей для своего обучения малых объемов априорной информации. В связи с этим нами предлагается новый подход, который мы назвали нечеткая пространственная экстраполяция в задачах медицинской диагностики.

Исходной информацией для рассматриваемого подхода являются данные, представленные в виде таблицы «объект-свойство», при этом важно отметить, что часть данных является размеченной, а часть – нет. Как уже отмечалось, классифицировать размеченные данные на основе традиционных методов распознавания образов не представляется возможным в силу малых объемов обучающей выборки. В статье [10] предложен подход к нечеткой кластеризации-классификации на основе комбинированного метода самообучения-обучения самоорганизующейся карты (SOM-LVQ), однако этот подход опять таки требует достаточно больших объемов обучающей информации, хотя и может решать задачи диагностики в условиях перекрывающихся классов. Рассматриваемый подход применительно к задаче диагностики может быть реализован в виде следующих действий.

Пусть имеется выборка наблюдений $x(k)=(x_1(k),x_2(k),...,x_i(k),...,x_N(k))^T \in R^n$, здесь n – объем выборки в таблице «объект-свойство». Пусть в данной выборке имеется $N_A+N_B+...+N_L+...+N_M$ размеченных (с установленным диагнозом) и $Q=(N-(N_A+N_B+...+N_L+...+N_M))$ неразмеченных данных, при этом к диагнозу-классу A относится N_A наблюдений, к диагнозу-классу B – N_B наблюдений и т.д., то есть

$$N_A+N_B+...+N_L+...+N_M<N,\ \hat{k}=1,2,...,\sum_{L=A}^{M}N_L.$$

Количество неразмеченных данных равно $Q=N-\sum_{L=A}^{M}N_L$, то есть наблюдения $x(\tilde{k})$, $\tilde{k}=\sum_{L=A}^{M}N_L+1,...,N$ (или $\tilde{k}=1,...Q$) с тем или иным уровнем принадлежности должны быть отнесены к соответствующим классам-диагнозам A, B, … L, M.

На первом этапе проводится расчет Q расстояний, заданных в Манхэттенской метрике:

$$d(\tilde{k},\hat{k})=\sum_{i=1}^{n}\left|x_i(\tilde{k})-x_i(\hat{k})\right|,$$

после чего рассчитываются уровни принадлежностей к конкретным классам:

$$\mu_L(x(\tilde{k}))=\frac{d^{-1}(\tilde{k},\tilde{k}\in L,\hat{k})}{\sum_{L=A}^{M}N_L}\Big/\sum_{k=1}^{}d^{-1}(\tilde{k},\hat{k}),\ L=A,B,\ldots,M.$$

На следующем шаге рассчитываются медианы уровней принадлежностей для каждого класса $\bar{\mu}_L(x(\tilde{k}))$, после чего можно рассчитать уточненные уровни принадлежностей:

$$\bar{\mu}^*_L(x(\tilde{k}))=\frac{\bar{\mu}_L(x(\tilde{k}))}{\sum_{L=A}^{M}\bar{\mu}_L(x(\tilde{k}))},$$

отвечающие условиям единичного разбиения.

В итоге для неразмеченной части данных мы получаем принадлежности к M возможным диагнозам.

Данная процедура позволяет компенсировать внутрикластерный разброс и разное количество наблюдений в разных классах, использование их в качестве расстояний Манхэттенской метрики позволяет ослаблять влияние от помех и выбросов, то есть придает процедуре робастные свойства [11]. Основными достоинствами предлагаемого подхода являются простота численной реализации и возможность обработки информации в условиях, когда количество доступных наблюдений соизмеримо с размерностью анализируемых векторов признаков.

Использование метода нечеткой пространственной экстраполяции в условиях дефицита информации

Работа предложенного метода нечеткой пространственной экстраполяции данных была апробирована на данных репозитория Калифорнийского университета (UCI Machine Learning Repository) Breast Cancer Wisconsin (Original) [12], которая представляет собой размеченную выборку данных, состоящую из 699 пациентов, каждый из которых характеризуется 10 признаками. В данных содержится 2 класса, характеризующих тип опухоли: доброкачественная или злокачественная. После удаления данных с пропусками и признака класса была получена выборка из 683 пациентов, которые характеризуются 9 признаками, из которой была сформирована обучающая выборка, состоящая из 40 пациентов: 20 относящиеся к 1 классу и 20, относящихся к 2 классу. Далее вся оставшаяся выборка в онлайн-режиме была обработана с помощью предложенного метода нечеткой пространственной экстраполяции. В результате была получена классификация всей выборки данных с нечеткой степенью принадлежности к классам. Ошибка классификации составила 3,85%.

Следующим шагом была попытка сравнить работу разработанного метода нечеткой пространственной экстраполяции с нечеткой кластеризацией-классификацией на основе комбинированного метода самообучения-обучения самоорганизующейся карты (SOM-LVQ) [10]. Следует отметить, что метод LVQ показывает примерно такую же точность при обработке предложенной выборки, однако с вычислительной точки зрения предложенный метод нечеткой классификации намного проще и он способен работать на обучающих выборках, состоящих из минимального количества пациентов (по одному из каждого класса). При этом для качественного обучения SOM-LVQ необходимо количество выборок обучающего массива как минимум соизмеримое с количеством признаков их характеризующих.

Выводы

Разработанный подход позволяет подвергать обработке медицинские данные, состоящие из ограниченного объема информации. Предложенный метод нечеткой пространственной экстраполяции позволяет работать с обучающими выборками, состоящими из 1 человека, что является актуальной проблемой для задач медицинской диагностики.

Список литературы

1. Han J. Data Mining: Concepts and Techniques / J. Han, M. Kamber. – Amsterdam: Morgan Kaufman Publ. – 2006. – 743 p.

2. Jose Valente de Oliveira, Witold Pedrycz Advances in Fuzzy Clustering and its Applications. – John Wiley&Sons Ltd, 2007. – 454 p.

3. Bishop C.M. Neural Networks for Pattern Recognition / C.M. Bishop. – Oxford: Clarendon Press, 1995. – 482 p.

4. Haykin S. Neural Networks. A Comprehensive Foundation / S. Haykin. – Upper Saddle River, N.J.: Prentice Hall, Inc., 1999. – 842 p.

5. Ruthowski L. Computational Intelligence. Methods and Techniques / L. Ruthowski. – Berlin-Heidelberg: Springer-Verlag, 2008. – 514 p.

6. Bezdek J.C. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing / J.C. Bezdek, J. Keller, R. Krisnapuram, N.R. Pal. – Springer 2005. – 776 p.

7. Du K.-L. Neural Networks and Statistical Learning / K.-L. Du, M.N.S. Swami. – London: Springer-Verlag, 2014. – 824 p.

8. Jang, J.-S. R. ANFIS: Adaptive-network-based fuzzy inference systems / J.-S. R. Jang // IEEE Trans. Syst., Man, and Cybern. – 1993. – 23. – N3. – P. 665-685.

9. Растригин Л.А. Экстраполяционные методы проектирования и удаления / Л.А. Растригин, Ю.П. Пономарев. – М.: Машиностроение, 1986. – 120 с.

10. Bodyanskiy Ye. Self-organizing map and its learning in the fuzzy clustering-classification tasks / Ye. Bodyanskiy, P. Mulesa, O. Slipchenko, O. Vynokurova // Вісник Національного університету «Львівська політехніка». Комп'ютерні науки та інформаційні технології. – 2014. – №800. – С. 83-92.

11. Мудров В.И. Метод наименьших модулей / В.И. Мудров, В.Л. Кушко. – М.: Знание, 1971. – 64 с.

12. WIlliam H. Wolberg UCI Repository of machine learning databases. – URL: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29. CA: University of California, Department of Information and Computer Science, 1992.

Поступила в редколлегию 21.07.2015

Рецензент: д-р техн. наук, проф. Е.А. Винокурова, Харьковский национальный университет радиоэлектроники, Харьков.

НЕЧІТКА КЛАСИФІКАЦІЯ ДАНИХ МЕДИКО-БІОЛОГІЧНИХ ДОСЛІДЖЕНЬ ЗА УМОВ ДЕФІЦИТУ ІНФОРМАЦІЇ

І.Г. Перова, Є.В. Бодянський

У статті розглянуто підхід, що дозволяє піддавати процедурі нечіткої кластеризації-класифікації вибірки медичних даних, сильно обмежені за обсягом з допомогою методу нечіткої просторової екстраполяції. Запропонована процедура відноситься до напрямку Medical Data Mining і являє собою гібридну систему, що дозволяє вирішувати задачу діагностування різного роду захворювань в умовах обмеженої вибірки, повного або часткового перекриття класів, різної їх щільності, різного чисельного наповнення і вимагає для свого навчання малих обсягів апріорної інформації.

Ключові слова: *нечітка кластеризація-класифікація, дефіцит інформації, нечітка просторова екстраполяція.*

FUZZY CLASSIFICATION OF DATA FOR BIOMEDICAL RESEARCH IN THE SCARCE INFORMATION

I.G. Perova, Ye.V. Bodyanskiy

In this paper the approach of fuzzy clustering-classification of medical data sample, limited in its dimensionality using the method of fuzzy spatial extrapolation is considered. The proposed procedure refers to the direction of Medical Data Mining, and is a hybrid system that can solve the task of diagnosing various diseases in a limited sample, complete or partial overlapping of classes, their different densities, different numerical filling and requires for its training of small volumes of a priori information.

Keywords: *fuzzy clustering , classification, deficit of information , fuzzy spatial extrapolation.*