Improving the Classification of chronic diseases using The Naive Bayes algorithm

Dr.Hoyam Omer Ali Abdallah
Emirates College of Science and Technology

Dr.Awad hassanMohamed
Napata College

مستخلص:

يعد التصنيف اليدوي للمرض إلى فئات مختلفة بناءً على مناطق المرضى السكنية مهمة شاقة وقد يختلف اعتمادًا على السيناريو المدروس . لذلك تم استخدام خوارزمية التصنيف Naïve Bayesian القياسية بهدف تصنيف المرض بناءً على عدة خصائص تمثل مناطقهم السكنية واعدادهم فيها وللتنبؤ السريع السهلايضا في عمليةتصنيف المناطق الاكثرعرضة للاصابة بالمرض المزمن كما تكمن اهمية الدراسة في تقليل العمل اليدوي العادي .تم استخدام منهجية تركزعلى مشكلة توسيع نموذج بايز التقليدي الساذج لتصنيف البيانات غير المؤكدة. المشكلة الرئيسية في نموذج بايز الساذج هي تقدير الاحتمال الشرطي للفئة ، وتقدير كثافة النواة هو طريقة شائعة لذلكامتدت طريقة تقدير كثافة النواة للتعامل مع البيانات غير المؤكدة. هذا يقلل المشكلة إلى اعتبار التكاملات المزدوجة. بالنسبة لوظائف النواة المحددة وتوزيعات الاحتمالية ، يمكن تقييم التكامل المزدوج بشكل تحليلي لإعطاء صيغة مغلقة ، مما يسمح بخوارزمية فعالة تعتمد على الصيغةبشكل عام ، ومع ذلك ، لا يمكن تبسيط التكامل المزدوج في أشكال مغلقة. في هذه الحالة ، يتم اقتراح نهج قائم على العينة. كماتشير النتائج التجريبية الرائعة إلى أن طريقة التصنيف المقترحة يمكن أن تكون واعدة ويمكن تطبيقها في مكان آخر وتساعد في عملية التشخيص حسب منطقة المريض . تم استخدام خوارزمية Naïve baseوذلك للتحقق من صحة الطريقة المقترحة تجريبياً لتكون دقيقة بنسبة 90٪ مما يثبت كفاءتها.

الكلمات المفتاحية : التصنيف ، المرضى ، الأمراض المزمنة ، نايف بايز

Abstract:

Manual classification of disease into different classes based on residential patient areas is a tedious task and may vary depending on the scenario studied. Therefore, the standard Naïve Bayesian classification algorithm was used to classify the disease based on several characteristics that represent their residential areas and their numbers, and for quick and easy prediction also in the process of classifying the areas most susceptible to chronic disease, and the importance of the study lies in reducing the normal manual work. A methodology focused on the problem of extending the traditional naive Bayesian model was used to classify uncertain data. The main problem of the naive Bayes model is estimating the conditional probability of the class, and estimating the kernel density is a common method so the kernel density estimation method has been extended to deal with uncertain data.This reduces the problem to considering double integrals. For finite kernel functions and probability distributions, the double integral can be evaluated analytically to give a closed formula, allowing an efficient formula-dependent algorithm in general, however, double integral cannot be simplified in closed forms. In this case, a sample-based approach is proposed.The remarkable experimental results also indicate that the proposed classification method can be promising and can be applied elsewhere and help in the diagnosis process by patient area. The Naïve base algorithm was used to validate the proposed method experimentally to be 90% accurate, which proves its efficiency.

Keywords: Classification,patients,chronic diseases,NaïveBayes

Problem statement :

Difficulty dealing with small-scale data and forecasting multiple, overlapping categories .

It is difficult to classify the areas most susceptible to chronic diseases.

Aims of this paper :

Extending the traditional, naive Bayesian model to classify uncertain data

Easy and fast prediction of data set category

Improving the condition of categorical variables compared to numerical variables

Introduction:

Naive Bayes is a classifier using probability and statistical methods proposed by a British scientist, Revered Thomas Bayes [(10)].

Naive Bayes often works much better in many complex real-world situations than might be expected [11].

Naïve Bayes is a popular model in Machine Learning applications because of its simplicity in allowing all attributes to contribute to the final decision equally. This simplicity is equivalent to computational efficiency, which makes the Naïve Bayes technique attractive and suitable for various fields.

Naïve Bayes is a subset of Bayesian decision theory. It's called naive because the formulation makes some naïve assumptions. Python's text-processing abilities which split up a document into a vector are used. This can be used to classify text. Classifies may put into human-readable form.It is a popular classification method in addition to conditional independence, overfitting, and Bayesian methods. [12]

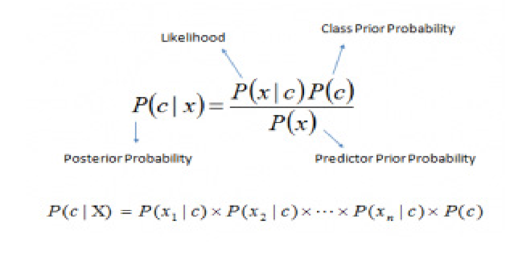
Considering a set of objects, each of which belongs to a known class, and each of which has a known vector of variables, the aim is to create a rule which enables to allocate future objects to a class, given just the vectors of variables marking out the future objects. These problems are known as ―supervised classification problem‖, are worldwide, and most of the methods for constructing such rules have been developed. It is very easy to establish, and no need any complicated repetitive parameter estimation schemes. This means it should be applied to huge data sets. It is easy to interpret, so unskilled users in classifier technology can make out the reason for it is making the classification it makes. Finally, it often does surprisingly well: it should not be the best possible classifier in any particular application, but it can usually be relied on to be robust and to do well.

ClassificationMethodology :

Naivebayesalgorithm

TheBayesTheoremformulaisasfollows:

It provides a method for calculating the suffix P (c | x) from P (c), P (x), and P (x | c). Look at the equation below:



Likelihood

Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

Posterior Probability

Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Where:

P(c | x) is the posterior probability of class (target) for the given predictor (attribute).

P(c) is the prior probability of class.

P(x | c) is the probability which is the probability of expecting a given class.

P(x) is the previous probability of the prediction.

Analysis and discussion

TrainingandTestDataSetSamples:

The following is the sample of the training group corresponding to the selection of the most common chronic disease in the eastern regions. The table clearly shows the different features used for the classification and the nature dependent on the application. The test data is also similar to the training data without class column which will be predicted with the help of the implementation of the algorithm described below.

Using the probabilities estimation below :

(chronic diseases = asthma , number of patients= 15 , areas of injury= East)

Table1 : Data set

Chronic diseases	Number of patients	Areas of injury	class
Heart diseases	17	The West	high
Stroke	15	The South	medium
Depression	17	The East	low
Asthma	17	The East	high
Lug cancer	25	The West	high
Stroke	18	The North	low
Asthma	15	The East	high
Asthma	22	TheEast	low

Although the number of crisis patients is large in the eastern region, the disease is classified as belonging to the LOW category because of the field of specialization that depends on the application.

Step 1 : We calculate the probabilities of the target attribute values :

P(high) = 40.5 = 8/

P(medium) = 10.1 = 8/

P(low) = 30.3 = 8/

Create probability table(2,3,4)(likelihood) :

Table 2:

Predicted expected	Low	high	medium
Heart diseases	0/3	1/4	0/1
Stroke	1/3	0/4	1/1
Depression	1/3	0/4	0/1
Lug cancer	0/3	1/4	0/1
asthma	1/3	2/4	0/1

Table 3:

Number of patients	Low	high	medium
15	0/3	1/4	1/1
17	1/3	2/4	0/1
22	1/3	0/4	0/1
25	0/3	1/4	0/1

Table 4:

Areas of injury	Low	high	medium
The West	0/3	2/4	0/1
The south	0/3	0/4	1/1
The north	1/3	0/4	0/1
The East	2/3	2/4	0/1

Classification result:

P(high\new instance)= P(high)*P(chronic diseases = asthma\ high)*P(number of patients= 15/high)*P(areas of injury=east/high)

P(high)= 0.5*(20.0312=(4/2*4/1* 4/

P(medium\new instance)= P(medium)*P(chronic diseases = asthma\ medium)*P(number of patients= 15/ medium)*P(areas of injury=east/ medium)

P(medium)= 0.1*(00=(1/0*1/1* 1/

P(low\new instance)= P(low)*P(chronic diseases = asthma\ low)*P(number of patients= 15/ low)*P(areas of injury=east/ low)

P(low)= 0.3*(10=(3/2*3/0* 3/

Then P(high) > P(low) & P(medium)

Results:

Study results based on previous experience

1. So the crisis is most prevalent in the east (Chronic diseases are categorized into different categories based on Relevant traits as the number of disease-related traits is increased to determine the original knowledge of the disease and categorized accordingly)
2. When the independence assumption is held, the naive

Bayes classifier performs better compared to other models such as logistic regression and requires less training data.

3. The above improved algorithm can be applied to any region of the world, where the algorithm has been improved by applying it to multiple samples of a data set andGet faster when training and query large numbers
4. Facilitate the diagnosis process according to the region.

Recommendations:

Using more than one data mining algorithm with multiple classification tasks to learn how to interact with properties and for further development and improvement.

Acknowledgments:

We would like to thank all those who encouraged or assisted them in this work.

Resources and references :

(1) H.Muhamadetal.,“Optimasi Naive Bayes Classifierdengan Menggunakan Particle Swarm Optimization pada Data Iris,” Teknol. Inf. Dan Pendidik., vol. 4, no. 3, pp. 180–184,2017.

(2) Arthur Chol, NazgolTavabi, Adnan Darwiche, ―Structured Features in Naive Bayes Classification‖, Association for the Advancement of Artificial Intelligence, 2016.

(3) WangS,JiangL,LiC.AdaptingnaiveBayestreefortextclassification. Knowledge and Information Systems. 2015;44(1):77–89

(4) J.Mañana-Rodríguez,“AcriticalreviewofSCImagoJournal&CountryRank,”Res.Eval.,pp.1–12,2014.

(5) S. Fitri, “Perbandingan Kinerja Algoritma Klasifikasi Naïve Bayesian , Lazy-Ibk , Zero-R ,DanDecisionTree-J48,”Dasi,vol .15,no.1,pp.33–37,2014.

(6) El Din Ahmed AB, Elarab IS. Data Mining: A prediction forStudent’sPerformanceusingClassificationMethod.WorldJournal of Computer Application and Technology.2014;2(2):43–7.

(7) T. R. Patil, “Performance Analysis of Naive Bayes and J48 Classification Algorithm forData Classification,” Int. J. Comput. Sci. Appl. ISSN 09741011-, vol. 6, no. 2, pp. 256–261,2013.

(8) J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, Third. Waltham,USA:ElsevierInc.,2012.

(9) G. Dimitoglou, J. a Adams, and C. M. Jim, “Comparison of the C4.5 and a Naive BayesClassifierforthePredictionofLungCancerSurvivability,”J.NeuralComput.,vol.4,no.8,pp.1–9,2012.

(10) I. Journal, S. Engineering, S. C. Applications, S. C. View, and M. Learning-based, “Is Na-ïve Bayes a Good Classifier for Document Classification ?,” Int. J. Softw. Eng. Its Appl.,vol.5,no. January,pp.37–46,2011

(11) Faisal KM, Mofizur RC, Alamgir H, Kesav D. Enhancedclassification accuracy on naive bayes data mining mod-els. International Journal of Computer Applications. 2011;28(3):9–16.

(12) Dong T, Shang W, Zhu H. An improved algorithm ofBayesian textcategorization.JournalofSoftware.2011;6(9):1837–43

(13) Toon Calders, SiccoVerwer, ―Three naive Bayes approaches for discrimination-free classification‖, Data Min Knowl Disk, 2010

(14) R. Entezari-Maleki, R. Arash, and M. Behrouz, “Comparison of Classification MethodsBased on the Type of Attributes and Sample Size,” J. Converg. Inf. Technol., vol. 4, no. 3,pp.94–102,2009.

(15) D. Xhemali, C. J. Hinde, and R. G. Stone, “Naive Bayes vs. Decision Trees vs. NeuralNetworks in the Classification of Training Web Pages,” Int. J. Comput. Sci., vol. 4, no. 1,pp.16–23,2009.

(16) HanJ,KamberM.DataMiningConceptsandTechniques.2nd Ed.2006.

(17) I. Rish, “An empirical study of the naive Bayes classifier,” Empir. methods Artif. Intell.Work.IJCAI,vol.22230,no. January2001,pp.41–46,2001.