دراسة إحصائية حول تحليل المكونات الرئيسية لنماذج المخلوط

إعداد

ندى عائض فهد القحطاني

رسالة مقدمة لنيل درجة الماجستير في العلوم (الإحصاء)

إشراف

د. زكية إبراهيم كلنتن

قسم الإحصاء , كلية العلوم

جامعة الملك عبد العزيز

جدة-المملكة العربية السعودية

١٤٤١هـ - ٢٠٢٠م

المستخلص

يستخدم علماء البيانات خوارزميات مختلفة للتعلم الآلي للعثور على أنماط في البيانات الضخمة التي تؤدي إلى رؤى عملية. لمعالجة هذه البيانات بشكل صحيح ، نحتاج إلى فحص ما إذا كان يمكن تفسيرها في مساحة منخفضة الأبعاد أم لا. في هذا البحث ، نستخدم تحليل المكونات الرئيسية (PCA) لتمثيل البيانات من الفضاء ذي الأبعاد العالية إلى الفضاء ذو الأبعاد المنخفضة والتعبير عن البيانات بهذه الطريقة لتسليط الضوء على أوجه التشابه والاختلاف بينهما. ثم اقترحنا سيناريوهين: الأول يتعامل مع البيانات المخفضة كنموذج خليط Gaussian واحد. بعد ذلك ، نحصل على تقديرات المعالم باستخدام خوارزمية التوقع إلى أقصى حد (EM). يتم تطبيق طريقة التجميع على بيانات مخفضة ، ثم تلائم نموذج الخليط في البيانات الجديدة عن طريق أخذ المتوسطات الطبقية كقيم أولية للمتوسطات لوسائل نموذج الخليط. السيناريو الثاني هو التعامل مع كل متغير في البيانات المخفضة بشكل فردي ، مرة عن طريق تركيب نموذج خليط Gaussian على كل متغير ، والمرة الأخرى عن طريق تركيب نموذج خليط Cauchy على كل متغير أيضًا. تتجلى فائدة استخدام نموذج خليط Cauchy في قدرته على التعامل مع البيانات غير المتجانسة وتلك ذات القيم المتطرفة. تم تقدير معلمات النموذج بناءً على خوارزمية تعظيم التوقعات (EM) .أثبتت فعالية الأساليب التي تمت مناقشتها من خلال دراسة المحاكاة ومجموعات البيانات الحقيقية.

في هذا البحث ، ناقشنا أيضًا تحليل المكونات الرئيسية للبيانات المختلطة (PCAMIX) وأظهرنا مدى فائدتها في معالجة البيانات الواقعية. في الوقت الحاضر ، معظم قواعد البيانات هي بيانات مختلطة ، مما يعني أن هناك مجموعة من المتغيرات الكمية والفئوية في قاعدة البيانات. يتم استخدام طريقة PCAMIX للتعامل مع هذا النوع من قواعد البيانات والسماح بجمع المعلومات الإحصائية على المجتمع المدروس. يتم التحقق من كفاءة PCAMIX باستخدام البيانات المتوفرة في حزمة R بالإضافة إلى بيانات المحاكاة.

A STATISTICAL STUDY ABOUT PRINCIPAL COMPONENTS ANALYSIS OF MIXTURE MODELS

By

Nada Ayed Fahd Alqahtani

A thesis submitted for the requirements of the degree of
Master of Science in Statistics

Supervised by

Dr. Zakiah Ibrahim Kalantan

DEPARTMENT OF STATISTICS
FACULTY OF SCIENCES
KING ABDULAZIZ UNIVERSITY
JEDDAH, SAUDI ARABIA
1441 H –2020 G

ABSTRACT

Data scientists use various algorithms of machine learning to find patterns in large data that lead to practical insights. To treat this data properly, we need to examine if it can be interpreted in a low-dimensional space or not. In addition, we try fitting the new data with different mixture models to obtain the suitable model. This step will perform the statistical model that predicts and estimates the parameters as close as possible to the original data. In this research, we use principal component analysis as a representation of the data from high dimensional to low dimensional space and expressing the data in such a way to highlight their similarities and differences. we proposed two scenarios: The first one is dealing with the reduced data as one Gaussian mixture model. Then, we obtain the estimations of the parameters by using the expectation-maximization algorithm. The clustering method is applied on reduced data, then fit the mixture model on the new data by taking the cluster means as initial values of the means for mixture model. The second scenario is dealing with each variable in the reduced data individually, once by fitting Gaussian mixture model on each variable, and the other time by fitting Cauchy mixture model on each variable also. The benefit of using the Cauchy mixture model is demonstrated in its ability to handle with heterogeneity and outliers. The model's parameters were estimated based on the expectation maximization algorithm. The effectiveness of the discussed methods demonstrated through a simulation study and by real datasets.

In this research, we also discussed the principal components analysis of mixed data (PCAMIX) and demonstrated how it is useful in today's real-world data. Nowadays, most databases are mixed data, meaning that there is a combination of numerical and categorical variables in the database. The PCAMIX method is used to handle this type of database and to allow statistical information to be collected over the studied population. The efficiency of PCAMIX is investigated using data set available in the R package and using simulated data.