دانشگاه علوم پزشکی کرمان

دانشکده بهداشت

پایان نامه مقطع کارشناسی ارشد رشته آمار زیستی

عنوان:

مقایسه ی خوشه بندی EA K- modes و NBEA K- modes یک روش جدید جهت خوشه بندی داده های گسسته و استفاده ی آن بر روی داده های مراقبت رفتاری و سرولوژیک در مصرف کنندگان تزریقی مواد

توسط: زهرا زمانی نسب

استاد راهنما: دکتر عباس بهرامپور

استاد مشاور: دکتر حمید شریفی

سال تحصیلی: ۱۳۹۶-۱۳۹۵

Kerman University of Medical Sciences

Faculty of Health

In Partial Fulfillment of the Requirements for the Degree of Masters in

Biostatistics

Title:

Comparing of EA K-modes Clustering and NBEA K-modes Clustering, A New Method for Clustering Categorical Data and Applying them on the Injecting drug Users Data Set

By:

Zahra Zamaninasab

Supervisor:

Abbas Bahrampour

Advisor:

Hamid Sharifi

Year:

2017

چکیده

مقدمه و اهداف: خوشه بندی یکی از روش های گروه بندی داده هاست، بطوریکه داده های شبیه به هم در یک خوشه قرار می گیرند. راه های زیادی برای خوشه بندی داده های پیوسته و رسته ای وجود دارد که در این بین خوشه بندی داده های رسته ای مقداری پیچیده تر است. یکی از راه های خوشه بندی داده های رسته ای، خوشه بندی K-مد است. در این مطالعه، ابتدا به بهبود بخشیدن خوشه بندی K-مد می پردازیم و با روش جمع اوری شواهد (Evidence Accumulation) بردار مدهای اولیه را ثابت می کنیم تا نتایج خوشه بندی K-مد با اجرای مجدد تغییر نکند. طبقه بندی ساده ی بیزی روش کلاس بندی برای پیش بینی کلاس های واقعی در یک مجموعه داده است. لذا با ترکیب دو روش EA K-modes و روش طبقه بندی ساده ی بیزی به روش جدیدی برای خوشه بندی داده های رسته ای منجر گردید که به نامNBEA K-modes نام گذاری و نتایج دو روش را باهم مقایسه خواهیم نمود.

روش اجرا: روش کار این مطالعه ابتدا بر روی پنج مجموعه داده از منبع داده های یادگیری ماشینیUCI، اجرا شد تا اعتبار روش EA K-modes برای تشخیص کلاس های واقعی موجود در داده ها سنجیده شود. برای اجرای خوشه بندی K-مد از نرم افزار Rو بسته ی klaR استفاده شد و الگوریتم های مربوط به روش جمع آوری شواهد در این نرم افزار نوشته شد. سپس از بسته ی e1071 برای خوشه بندی به روش طبقه بندی ساده ی بیز، برای خوشه بندی داده های ملی مصرف کنندگان تزریقی مواد، استفاده گردید.

یافته ها: ابتدا روش EA K-modes را بر روی پنج مجموعه داده ی واقعی اجرا گردید تا بردار مد های ثابت اولیه بدست آید، سپس خوشه بندی K-مد با استفاده از همان بردار مد اولیه ی ثابت اجرا شد. نتایج برای هر پنج مجموعه داده نشان داد که نرخ خلوص خوشه ها در روش EA K-modes نسبت به K-مد سنتی افزایش قابل توجهی پیدا کرده است. سپس روش EA K-modes را روی داده های مصرف کنندگان تزریقی مواد(IDU) اجرا و در نهایت با ترکیب طبقه بندی ساده ی بیز و روش EA K-modes به روش جدیدی به نام NBEA K-modes منجر گردید و برای خوشه بندی داده های IDU از این روش نیز استفاده شد. نتایج با تعداد دو؛ سه؛ چهار و پنج خوشه مورد مقایسه قرار گرفت.

بحث و نتیجه گیری: نتایج خوشه بندی به روش NBEA K-modes در مقایسه با روش EA K-modes در تمام تعداد خوشه های متفاوت به افتراق بهتری رسید و بهترین تعداد خوشه ها برای این مجموعه داده، سه خوشه در نظر گرفته شد. نتایج مطالعه ی ما با نتایج مطالعه ی Aranganayagi و همچنین نتایج مطالعه ی Khan، هم خوانی داشت.

کلید واژه ها: خوشه بندی، K-modes، جمع آوری شواهد، طبقه بندی ساده ی بیز، داده های رسته ای

Abstract

Background: Clustering is the method of grouping subjects, those who are similar together stay in the same cluster. There are many ways for clustering data with continuous or discrete variables. Among these methods, clustering of discrete data is more complicated. One of the methods for clustering discrete data is the K-modes method. In this research we improve the K-modes results with Evidence Accumulation (EA) method that helps to fix the initial mode vector, then we apply the Naïve Bayes method on combination of K-Mode and EA. Naïve Bayes classifier is the classification method to predict the unknown real classes. Finally the results of EA K-Mode, NBEA K-modes will be compared.

Method: The methods are applied on five real datasets of the UCI Machine Learning Data Repository for checking the external validity and purity of our methods which the true classes are determined. The software R with package klaR for K-modes, EA and package e1071 for Naïve Bayes are used. Also the methods are applied on Injecting Drug Users (IDU) national data set with 2546 subjects and 22 independent variables.

Results: We applied EA K-modes algorithm on five real datasets and fixed the initial modes, then we run again the K-modes algorithm with the fixed initial mode vector. The results indicate that the purity rate in the EA K-modes has significant improvement in comparison of classic K-modes algorithm in all of the five datasets. We use the results of combination method to run the EA K-modes on the Injecting Drug Users dataset that has no real class labels. Finally we applied the Naïve Bayes classifier with prior of cluster membership that find in EA K-modes. For K=2,3,4,5 the results indicate that EA K-modes with Naïve Bayes concept made better separation among the subjects and better clustering in comparison with mentioned methods.

Discussion and Conclusion: In this paper, we proposed a new method for clustering for binary data and it is Naïve Bayes EA K-modes. The results showed that our new method leads to stable clustering results compare with the study of Aranganayagi that was used classic k-modes only and the study of Shehroz S Khan that was used the EA K-modes method only. The Naïve Bayes EA K-modes method improves the purity and make better separation in data set.

Keywords: clustering, K-modes, Evidence Accumulation, Naïve Bayes classifier, discrete data