Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2018; 4(2): 26-32

Original Article

Naïve Bayes evidence accumulation K-modes clustering: A new method for classifying binary data and its application on real data of injecting drug users

Zahra Zamaninasab[1], Hamid Sharifi[2], Abbas Bahrampour[3]*

[1] HIV/STI Surveillance Research Center, and WHO Collaborating Center for HIV Surveillance, Institute for Futures Studies in Health, Department of Biostatistics and Epidemiology, School of Public Health, Kerman university of Medical Sciences, Kerman, Iran
[2] HIV/STI Surveillance Research Center, and WHO Collaborating Center for HIV Surveillance, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran AND Department of Biostatistics and Epidemiology, Faculty of Public Health, Kerman University of Medical Sciences, Kerman, Iran
[3] Modelling in Health Research Center, Institute for Futures Studies in Health, Department of Biostatistics and Epidemiology, Health Faculty, Kerman University of Medical Sciences, Kerman, Iran

ARTICLE INFO

Received 17.07.2017
Revised 19.08.2017
Accepted 01.11.2017

Key words:
Clustering;
Evidence accumulation;
Naïve Bayes classifier;
Discrete

ABSTRACT

Background & Aim: Clustering is the method of classifying discrete data such as K-modes, and Naïve Bayes classifier is the classification to predict the unknown real classes. In this research, we improve the K-modes results by applying the Evidence Accumulation (EA) method to keep the initial mode vector to use in the Naïve Bayes EA K-Mode.
Methods & Materials: The methods are applied to four real datasets, which the true classes are specified, for checking the external validity and purity of our methods. The free programming software R with package klaR for K-modes, EA, and package e1071 for Naïve Bayes is used. In addition, the methods are applied to the data of Injecting Drug Users (IDU) national dataset with sample size 2546.
Results: The EA K-modes algorithm applied to five real datasets then with the kept initial mode vector, rerun the K-modes. The results indicate the purity in the EA K-modes (0.544, 0.862, 0.914, 0.944, 0.625) has significant different with classic K-modes (0.497, 0.610, 0.404, 0.650, 0.625). Finally, we applied the Naïve Bayes classifier with prior probability finds in EA K-modes. For K=2 Naïve Bayes EA K-modes made better clustering (0.71, 0.873 against 0.625, 0.862 EA k-mode and 0.497, 0.61 K-mode).
Conclusion: In this paper, we proposed Naïve Bayes EA K-modes as a new method for clustering of binary data. Our new method leads to stable clustering compare with the previous studies. The Naïve Bayes EA K-modes method improves the purity and establishes a better separation.

Introduction

Clustering is one of the most important and useful methods in data mining process which uses for recognizing similarities in a data set. In this method, a given data set is partitioned into k clusters such that the subjects in each cluster are more similar to each other than the subjects in different clusters. The main goal of clustering analysis is to detect whether or not the data is heterogeneous, and whether the data fall into distinct clusters. Grouping data is provided by using some similarity or dissimilarity measures, such as Euclidean distance (1-3). K-means clustering method is one of the most useful algorithms for classifying continuous data, but in some cases, the data contains discrete variables. The geometric properties and distance measures such as Euclidean distance cannot be used for the discrete data (4). The clustering of discrete data is more complicated than continuous type.

Huang in 1998 was proposed an algorithm for clustering categorical data by using a simple matching dissimilarity measure called k-modes, which is an extension of k-means and uses the modes, instead of means for clusters centroids (5). Most of the algorithms for clustering discrete data, like k-modes require a selection of

* Corresponding Author: Abbas Bahrampour, Postal Address: Modelling in Health Research Center, Institute for Futures Studies in Health, Department of Biostatistics and Epidemiology, Health Faculty, Kerman University of Medical Sciences, Kerman, Iran.
Email: abahrampour@yahoo.com

Please cite this article in press as: ZamaninasabZ, Sharifi H, Bahrampour A. Naïve bayes evidence accumulation K-modes clustering: A new method for classifying binary data and its application on real data of injecting drug users. *J Biostat Epidemiol. 2018; 4(2): 26-32*

k randomized initial points, which leads to a problem that clusters depend on initial points and so number of iteration which also leads to different clustering. Furthermore, inappropriate selection of initial modes leads to undesirable clustering results (6).

One method to control part of the problem in k-modes is clustering with evidence accumulation (EA) method, which it stores the results of multiple clustering in a mode-pool and then combine them into a single data mode. So we extended this method in two ways, first of all we run k-modes clustering for N times and secondary combine the results with EA method and Naïve Bayes classifier to achieve the stable mode vector. With this mode vector, the k-modes clustering will run once again, and then use the cluster membership vector of k-modes clustering as a prior for Naïve Bayes classifier method. In Naïve Bayes classifier, each object assigns to the group with maximum posterior probability (3). With keeping the vector of modes and hence the stability of clusters we expected that the results of Naïve Bayes EA K-modes clustering are more close to reality than classic k-modes clustering methods.

In the research that was conducted by Aranganayagi et al, the k-modes clustering and proposed method that was K-modes clustering with Naïve Bayes concept run on some real datasets such as Balance Scale, Congressional Votes, Soybean small, and Breast cancer datasets. It seems that in the proposed method the repetition execution and updating of modes don't be needed and it was expected that the proposed method is more efficient than k-modes and makes better purity rates (7). In a study by Shehroz S Khan et al, the evidence accumulation process was used for getting stable initial modes and tested on the real datasets, which are mentioned. After using EA method combine with k-modes clustering, the results were similar to the actual modes of the datasets. Therefore, the clustering process achieved faster convergent results (6).

In this study, we first use five real data sets of the UCI Machine Learning Data Repository and show the results of using the EA K-modes method achieve a stable mode vector, instead of classic k-modes. Then we apply the EA K-modes and Naïve Bayes EA K-modes (the new method) for the clustering the two data sets of UCI and the national data set of people who Inject Drugs in Iran (IDU). So with this new method, it is expected to cover the gaps of previous methods.

Our objective of study other than introducing the new method is to apply the method to the data of injected drug users for clustering them, to identify and specify the groups which consist of people with common behaviors and planning to prevent such behaviors that lead them to drugs. Because of all the variables in this dataset are discrete, we must use the K-modes method among all clustering methods. The classic K-modes method leads to unstable results and different clusters in each running of K-modes method that is unfavorable. Therefore, we try to determine the stable results with combining of EA K-modes and Naïve Bayes method, instead of classic K-modes.

Methods

K-modes Algorithm for Clustering Discrete Data

K-modes clustering algorithm which is based on K-means paradigm with some extensions like using a simple matching dissimilarity measure (Hamming Distance) and modes instead of means, can control the limitation of K-means method and can be applied for clustering of discrete data. Simple dissimilarity measure of k-modes clustering is defined as follow:

Let X and Y and are two objects with variables. The dissimilarity measure between X and Y, or d (X,Y) and, or is defined by the total mismatches of corresponding variable categories of this two objects. The smaller number of mismatching corresponding variables, more similarity of two objects, which mathematically can be written as follow:

$$d(X,Y)=\sum_{j=1}^{F}\delta(x_j,y_j)$$

Where the $\delta(x_j.y_j)$ is defined by:

$$\delta(x_j.y_j)=\begin{cases}0 & (x_j=y_j)\\ 1 & (x_j\neq y_j)\end{cases}$$

If $Z=\{Z_1,Z_2,\ldots,Z_n\}$ is a set of categorical data objects, each object describes with categorical variables, a mode of is a vector like $Q=\{q_1,q_2,\ldots,q_F\}$ which minimized the

following function:

$$D(Z,Q)=\sum_{i=1}^{n} d(Z_i,Q)$$

The algorithm of K-modes has four steps as follows:

1. Selection of k initial modes
2. Assign object to the cluster whose mode is nearest to it according to equation d(X.Y).
3. Compute the new modes of each cluster
4. Repeat steps 2 and 3 until no object has changed its cluster membership.

Computing initial modes using EA

In general, there are several ways to accumulate evidence:

a) Combining the results of different clustering methods.

b) Compute different results by bootstrapping and other re-sampling methods.

c) Running a given clustering method for N times with different initial points and combine the results.

In this research, we follow the step (c) for accumulating the evidence by running the k-modes algorithm with different initial modes N times and repeat the process until reaching a convergent mode vector. In addition, we store the convergent mode vector results of k-modes in a mode-pool then we combine the results with this following algorithm in R to calculate a fixed initial mode:

This algorithm in R software can be written as follows (for known N and K):

```
Pool_mode=list(0)
for (i in 1:N){
cluster1=kmodes(data,K,
iter.max=10,weighted=FALSE)
m=as.matrix(cluster1$modes)
mode[i]=list(m)
}
Pool_mode
```

Combining the results and extracting the initial modes from pool-mode

From the above algorithm we compute the pool-mode that is a list of N, K*F modes. Recall that the F is the number of variables in the data set. Then we extract the most frequent modes which can be used as fixed initial mode with following algorithm in R: (for known k, N, and for a given F)

```
initial=matrix(NA,K,F)
v=c(NA)
for (i in 1:K){
for (j in 1:F){
for (k in 1:N){
v[k]=Pool_mode[[k]][i,j]
fr=table(v[k])
mod=as.numeric (names(fr[fr==max(fr)]))
initial[i,j]=mod
}}}
Initial
```

After determining the convergent pool-mode, we have only one unique initial mode vector which leads to invariant clusters. After running the above algorithm, we have a K*F matrix of initial modes.

Naive Bayes Classifier

Naïve Bayes method is a probabilistic model to classifying the data. This method assumes that the predictor variables are independent. Let $Z=\{Z_1,Z_2,...,Z_n\}$ is a set of discrete data, each object F has variables and the number of real classes in the data is K, $C=\{C_1,C_2,...,C_k\}$ Naïve Bayes method assigns the object Z to the class C_i, if

$$P(C_i|Z)>P(C_j|Z) \quad 1\le i,j\le K$$

From the Bayes theorem, the posterior probability for each object in each K classes is defined as bellow:

$$P(C_i|Z)=\frac{P(Z|C_i)P(C_i)}{P(Z)}$$

In the above equation,

$$P(Z|C_i)=P(z_1|C_i)*P(z_2|C_i)*...*P(z_F|C_i)$$

In the last formula each part of $P(Z_f|C_i)$ is a fraction. The numerator shows the number of objects in the class i that its fifth value is equal to z_f and denominator shows the total number of objects in class i.

Naïve Bayes EA K-modes clustering

In our method, after determining the fixed mode vector from the convergent process of evidence accumulation method, we run the k-modes clustering with this mode vector and use the cluster membership vector of its result as a prior distribution for Naïve Bayes method. The dataset file is in SPSS format and we replace the vector of cluster membership of EA k-modes instead of class column, and run the Naïve Bayes

model. First, we apply this new combined method on standard data, which their true classes are specified, and then we use the method on data of IDU data. Therefore, this study extends the previous methods in two ways:

First: Fixing the initial modes

Second: Applying the fixed vector of initial modes as priors of Naïve Bayes model to calculate the posteriors.

Results

Experimental Results

We test our approach, Naïve Bayes EA K-modes clustering on the real data sets of UCI Machine Learning Data Repository where the true classes are specified:

A) Balance Scale: The balance scale data set consist of 625 objects with five variables. Each object is shown with three classes:

B (Balanced), L (Left), R (Right)

B) Congressional Votes: There are 435 objects with 16 variables. All these variables are binary (yes or no) and each object are labeled with two classes:

R (Republican), D (Democrat)

C) Soybean small: This is the data set of 47 objects with 35 variables, but only 21 of variables are appropriate for our experiment.

One of the four diseases labels each object.

D) Breast cancer: The breast cancer data set contains 699 objects with 11 variables. Each object is labeled with two classes:

Two (Benign), four (malign).

E) Lenses: This is the data set of 24 objects with four categorical variables. Each object is labeled with two classes:

One (with fitting lenses), two (without fitting lenses)

In Table 1, we list the results of purity rates of five data, which shows that the k-modes clustering with evidence accumulation concept is more efficient than k-modes.

Furthermore, in the congressional votes dataset and Lenses data set in Tables 2 and 3, after running the k-modes with EA concept, the results are assumed as a prior information for

Table 1. Comparison of purity rate results of k-modes and k-modes with EA process

Data set	#Clusters	K-modes	EA k-modes
Balance Scale	3	0.497	0.544
Congressional Votes	2	0.610	0.862
Soybean small	4	0.404	0.914
Breast cancer	2	0.650	0.944
Lenses	2	0.625	0.625

Table 2. Classifying congressional votes dataset of the real data by the EA K-mode and Naive Bayesian EA k-mode and the Purities

Cluster/Class	EA k-mode 1	Naive Bayesian EA k-mode 1	EA k-mode 2	Naive Bayesian EA k-mode 2	TOT	EA k-mode	Naive Bayesian EA k-mode	EA k-mode purity	Naive Bayesian EA k-mode purity
1	14	8	153	159	167	153	159	0.862	0.873
2	221	220	46	47	267	221	220		
Total	235	228	199	207	434	374	379		

Table 3. Classifying lenses dataset of the real data by the EA K-mode and Naive Bayesian EA k-mode and the Purities

Cluster/ Class	EA k-mode 1	Naive Bayesian EA k-mode 1	EA k-mode 2	Naive Bayesian EA k-mode 2	TOT	EA k-mode	Naive Bayesian EA k-mode	EA k-mode purity	Naive Bayesian EA k-mode purity
1	5	5	4	4	9	4	4	0.625	0.71
2	7	13	8	2	15	7	13		
total	12	18	12	6	24	11	17		

Table 4. Characteristics of 22 independent variables

Variable Name	0 (%)	1 (%)	Description
Sex	2.6	97.4	0: female, 1: male
Age (year)	60.4	39.6	0: < 35, 1: >= 35
Marriage status	69.3	30.7	0: single, 1; married
Education	70.5	29.5	0: under diploma, 1: diploma and higher
Job	93	7	0: non-constant, 1:constant
Salary	40.5	59.5	0: without income, 1: with income
Age of first drug use	27.5	72.5	0: <=15, 1: >15
Age of first injection	2.8	97.2	0: <=15, 1: >15
Injecting drug in last month	37.2	62.8	0: no, 1: yes
Time of last injecting	53.7	46.3	0: in last week, 1: before last week
Risk level of used syringe	96.8	3.2	0: low risk, 1: high risk
history of prison	22.2	77.8	0: no, 1: yes
Sexual	16.4	83.6	0: no, 1: yes
Sex with a non-monetary partner last year	75.5	24.5	0: no, 1: yes
Sex with a monetary partner last year	81.1	18.9	0: no, 1: yes
Men with men sex	85.9	14.1	0: no, 1: yes
Knowledge of HIV	3.7	96.3	0: no, 1: yes
Self at risk for HIV	39.9	60.1	0: no, 1: yes
HIV test	46.7	53.3	0: no, 1: yes
Aware of result of last HIV test	23.1	76.9	0: no, 1: yes
Sexually Transmitted Infections (STI)	90.6	9.4	0: no, 1: yes
Knowledge about Transmission of HIV	29.8	70.2	0: low knowledge, 1: high knowledge

Naïve Bayes model. We replace the cluster membership of EA k-modes instead of real classes of congressional votes and Lenses data in the data set, which shows the new method, improve the purity.

Table 4 shows the characteristics of all independent variables in IDU data set:

Running the proposed method on IDU dataset

As it is showed that, the Naive Bayes EA K-modes has better purity in comparing with the previous methods so we applied this method clustering to IDU data. This Behavioral and serological surveillance data set was collected with questionnaire in 17 provinces of Iran and Consists of 2546 subjects who are injected drugs.

First, the EA k-modes algorithm used to achieve the fixed mode vector, for k=2.

Table 5 showed in both methods variables injecting drug in last month, self at risk for HIV,

Table 5. Comparison of clustering results with two methods in cluster 1

Cluster 1									
EA K-modes					**Naïve Bayes EA K-modes**				
Variables	0	1	0 (%)	1(%)	Variables	0	1	0(%)	1(%)
Injecting drug in last month	200	1559	21.00	97.50	Injecting drug in last month	1	1550	.10	97.00
Self at risk for HIV	483	1276	47.50	83.40	Self at risk for HIV	499	1052	49.11	68.70
Sex with a monetary partner last year	1412	347	68.40	72.14	history of prison	303	1248	53.50	63.00
Sex with a non-monetary partner last year	1321	438	68.70	70.30	Salary	602	949	58.40	62.60
Salary	696	1063	67.50	70.00	HIV test	706	845	59.40	62.20
Age of first drug use	469	1290	67.00	69.80	Sex	13	1538	19.70	62.00
Sex	31	1728	46.96	69.67	Age of first drug use	407	1144	58.00	62.00
Age of first injection	43	1716	60.60	69.30	Age of first injection	28	1523	39.40	61.53
Knowledge of HIV	62	1697	65.20	69.20	Knowledge about Transmission of HIV	452	1099	59.50	61.50
Knowledge about Transmission of HIV	527	1232	69.40	69.00	Sex with a monetary partner last year	1256	295	60.80	61.30
Sexual	295	1464	70.70	68.70	Knowledge of HIV	52	1499	54.70	61.10
Men with men sex	255	1504	71.30	68.70	Sexual	260	1291	62.30	60.60
history of prison	374	1385	66.00	68.60	Men with men sex	225	1326	62.80	60.60
Aware of result of last HIV test	433	1326	73.50	67.70	Aware of result of last HIV test	366	1185	62.10	60.50

Table 5. Cntd

STI	1599	160	69.30	66.90	Sex with a non-monetary partner last year	1175	376	61.10	60.35
Education	1261	498	70.30	66.20	STI	1413	138	61.20	57.70
Age	1109	650	72.00	64.40	Education	1118	433	62.30	57.60
Marriage status	1264	495	71.61	63.40	Age	992	559	64.50	55.40
HIV test	927	832	78.00	61.30	Marriage status	1126	425	63.80	54.40
Risk level of used syringe	1709	50	69.30	61.00	Job	1477	74	62.40	41.60
Job	1661	98	70.14	55.00	Risk level of used syringe	1531	20	62.00	24.40
Time of last injecting	1368	391	100.0	33.00	Time of last injection	1361	190	99.50	16.00

Table 6. Comparison of clustering results with two methods in cluster 2

Cluster 2									
	EA K-modes					**Naïve Bayes EA K-modes**			
Variables	0	1	0 (%)	1(%)	Variables	0	1	0(%)	1(%)
Time of last injection	0	787	.00	66.80	Time of last injection	7	988	.50	83.80
Job	707	80	29.80	44.90	Risk level of used syringe	933	62	1.40	75.60
Risk level of used syringe	755	32	30.60	39.00	Job	891	104	37.60	58.40
HIV test	261	526	22.00	38.70	Marriage status	639	356	36.20	45.60
Marriage status	501	286	28.40	36.60	Age	545	450	35.40	44.60
Age	428	359	27.80	35.60	Education	676	319	37.70	42.40
Education	533	254	29.70	33.70	STI	894	101	38.70	42.20
STI	708	79	30.70	33.00	Sex with a non-monetary partner last year	748	247	38.90	39.60
Aware of result of last HIV test	156	631	26.50	32.20	Men with men sex	133	862	37.20	39.40
Sexual	122	665	29.30	31.20	Aware of result of last HIV test	223	772	37.70	39.40
Men with men sex	103	684	28.70	31.20	sexual	157	838	37.60	39.30
Knowledge about Transmission of HIV	232	555	30.56	31.00	Knowledge of HIV	43	952	45.30	38.80
Knowledge of HIV	33	754	34.70	30.76	Sex with a monetary partner last year	809	186	39.20	38.70
Age of first injection	28	759	39.40	30.70	Age of first injection	43	952	60.60	38.50
Sex	35	752	53.00	30.32	Knowledge about Transmission of HIV	307	688	40.40	38.50
Age of first drug use	231	556	33.00	30.00	Sex	53	942	80.30	38.00
history of prison	192	595	33.90	30.00	Age of first drug use	293	702	41.80	38.00
Salary	335	452	32.50	29.80	HIV test	482	513	40.57	37.80
Sex with a non-monetary partner last year	602	185	31.30	29.70	Salary	429	566	41.60	37.40
Sex with a monetary partner last year	653	134	31.60	27.80	history of prison	263	732	46.50	37.00
Self at risk for HIV	533	254	52.50	16.60	Self at risk for HIV	517	478	50.80	31.20
Injecting drug in last month	748	39	78.90	2.40	Injecting drug in last month	947	48	99.90	3.00

time of last injecting, and risk level of used syringe are the most common variables used for built in cluster1. Table 6 showed in both methods, variables age of first drug use, age of first injection, Sexually Transmitted Infections (STI), and HIV test are the most common variables used for built in cluster 2.

Discussion

As the data mining deals with large datasets the clustering algorithm should be scalable. The K-modes clustering process is scalable but the repetitive execution is needed to minimize the D (Z.Q). The cause that in each execution of K-modes the clustering results was differ from each other. Therefore, we use an approach to compute the fixed initial modes for K-modes to clustering discrete data using evidence accumulation concept. We used the process of EA for combining the results of multiple K-modes clustering and save them in a pool-mode.

Therefore, in this study, first we run the classic K-modes and EA K-modes on four real datasets Balance scale, Congressional votes, Soybean small, and Breast cancer that the true classes are specified. With comparing the purity

rate of two methods we see that using the EA K-modes for achieve the fix initial mode vector get the higher purity rate. After determining the convergent pool-mode, we applied this method for clustering IDU dataset to fix the initial mode vector and because in IDU data we don't have the true class labels for each subject. But use of EA K-modes in real datasets says us that the clustering results was very closer to true classes so we applied Naïve Bayes concept on IDU data, using cluster membership of EA K-modes. We compare the results of two approaches EA K-modes and Bayesian EA K-modes.

In the research of Shehroz S Khan et al (8), the Evidence Accumulation method was introduced to fix the initial mode vector, it shows that the modes computed from EA method were found to be similar to the actual modes of those datasets. The results of the section that using the EA method for clustering in the study are coherence with the results of mentioned study. Both studies are showed that clustering K-modes with EA concept makes better purity.

In congressional votes dataset, because of all variables are binary like variables in our study, we applied the EA K-modes and the Naïve Bayes EA K-modes. The results showed that the purity rate in Naïve Bayes EA K-modes is higher than the purity rate in EA K-modes. Research of Aranganayagi et al (7), showed that Naïve Bayes k-modes algorithm is efficient than classic K-modes and has high purity in clustering but the restriction of this study is that the modes are not fixed and so leads to different clustering in each repetition. In this study, we applied the Naïve Bayes EA K-modes and EA K-modes that using the EA method in clustering achieve the stable and convergent mode vector and so invariant clustering results.

Conclusion

In present study, the result show that the Naïve Bayes EA K-modes constructs better separation in datasets and leads to improve the purity and so fixed clustering.

Acknowledgements

The authors gratefully thank HIV/STI Surveillance Research Center. We profoundly thank the people who were generous with their time and participated in the study. This paper presents the results of Zahra Zamaninasab's thesis for Master of Science in Biostatistics.

Ethical Approval and Consent to participate

The ethics committee of Kerman University of Medical Sciences reviewed and approved the study's protocol. The ethical code is IR.Kmu.REC.1394.610

References

1. Guha S, Rastogi R, Shim K, editors. CURE: an efficient clustering algorithm for large databases. ACM Sigmod Record; 1998: ACM.
2. Berkhin P. A survey of clustering data mining techniques. Grouping multidimensional data: Springer; 2006. p. 25-71.
3. Han J, Pei J, Kamber M. Data mining: concepts and techniques: Elsevier; 2011.
4. Rencher AC. Methods of multivariate analysis: John Wiley & Sons; 2003.
5. Huang Z, editor. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. DMKD; 1997.
6. Khan SS, Kant S, editors. Computation of Initial Modes for K-modes Clustering Algorithm Using Evidence Accumulation. IJCAI; 2007.
7. Aranganayagi S, Thangavel K. Clustering categorical data using bayesian concept. Intl J Comput Theory Engin. 2009;1(2):119.