International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Clustering Algorithm Based on Local Random Walkwith Distance Measure

Gang Dai, Baomin Xu

School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044 China

Abstract: ***Cluster analysis is widely used in the field of data mining. However, the K-means algorithm which is widely used has a strong sensitivity for the initial values. Namely, the parameters such as clustering coefficient and centroid should be determined when the cluster is initialized. In the paper, we propose a K-means algorithm that based on link information and regard KL divergence distance as the objective function. This method not only introduces the way of the local random walk with the shortest path, but also uses the link information to convert the distance space. In other word, we utilize the local random walk with the shortest path to convert the distance between data into the transition probability of the random walk. Then, we use the random walk realize the conversion of the distance space. The core concept is the distance of converting node pair that refers to the node to the whole network node distance. The experimental results show that the proposed algorithm can improve the cluster result efficiently.***

Keywords: Random walk; K-means; Clustering; KL divergence; Complex system.

1. Introduction

The goal of clustering is to find the natural classification of the data set. Finding some kind of way to partition the data set makes the similarity of data in the same cluster higher than the different one. Because clustering is an unsupervised method, it can find pattern from data set. This makes clustering be widely used in many fields, such as information retrieval, Biocomputing, web analysis and so on. Clustering has been extensively used in the fields of pattern recognition, statistics and machine learning. Typically, the most common algorithms are hierarchical clustering algorithm [1-3] and K-means algorithm [4]. Meanwhile, the clustering algorithm based on probability model [5, 6] is becoming more and more widely applied.

In the data set which has link relations, there are many definitions for the cluster element object. We can not only cluster for single entity, but also for the set of the entity, even for the subgraph of original data set. The key to this kind of problem is how to define the similarity of two entities or two subgraphs via potential network structure. There is not too much research work for the clustering using subgraph, and the reference [7] offer the earlier research work.

The clustering based on link isalso widelyused, such as finding hub nodes via web set clustering to recognize the mirror site, finding the authors who always publish together via publishing field clustering or finding a new research field via the cluster which has the same invention and discovery. For example, clustering in the epidemic field can help to find the diseases which have the same contact people or the same propagation mode[8].

The concept of K-means appeared in 1967 [9]. However, its idea can be traced back to 1957[10]. K-means is an easy and fast classics clustering algorithm and it is fit to combine with other algorithm to resolve the practical problem. In the scientific research, K-means can be used to develop new algorithm and reduce complexity, therefore it is beneficial for researcher to focus on the effect of the new algorithm. This is also the reason of using K-means algorithm while using the random walk to cluster analysis. The major defect of K-means is the strong sensitivity to the initial value. So we must determine the clustering coefficient and centroid when initiate the cluster. Now, there are many researchers work at improving the defect of K-means, such as clustering analysis based on random walk [16].

This paper proposes a new K-means algorithm using shortest path and random walk. The new K-means clustering algorithm uses a different way from other algorithms based on K-means in using the link information of data points. This method converts the distance between data points into the transition probability of random walk, then, proceeds walking. In this way, it can realize the conversion of the distance space. Its essence is using the initial distance of data points to the other points on the whole network to compute the final distance.

2. Preliminaries

(1) The Model of Random walks based on shortest path

Let $G(V,E)$ be an acyclic undirected graph. $V$is the set of vertices and E is the set of edges. In LRW[11], walker runs to an arbitrarily neighbor node with probability $1/K$. $k$is the degree of the node. So we can get the adjacent matrix which is the one-step transition probability matrix$P$. $P(i,j)$ is the probability of a random walker starting at node i and moving to node j. If node I is directly connected to node j, the value of $P(i,j)$ is $1/k_i$, otherwise, the value is 0. We also give a vector $\vec{\pi}_i(t)$ which is the probability of a random walker starting in node i to reach node j after t steps. The initial transfer vector $\vec{\pi}_i(0)$ means that walker's initial probability is 1 at node i. The probability of transfer node is

$$\vec{\pi}_i(t) = P^T \vec{\pi}_i(t-1) \qquad (1)$$

LRW defines a similarity metric

$$s_{ij}^{LRW}(t)=\frac{k_i}{2|E|}\cdot\pi_i(t)+\frac{k_j}{2|E|}\cdot\pi_j(t) \quad (2)$$

$|E|$is the number of edge. We assume the shortest path of nodes is the steps of random walks. So it is not necessary to use uniform optimal steps for the whole network. Then we assume $f_{ij}^{(n)}=P(T_i=n)$to be the probability that the walker arriving at node j from node i for the first time after n steps. $\sum_{n=1}^{\infty}f_{ij}^{(n)}$is the final probability of a random walker starting from node i to node j. d_{ij} is the shortest path between node i and node j. Obviously, when $n<d_{ij}$, the value of $f_{ij}^{(n)}$ is 0. So the first-time-passage probabilityof nodes is $f_{ij}^{d_{ij}}$. Using the $f_{ij}^{d_{ij}}$ approximation $\sum_{n=1}^{\infty}f_{ij}^{(n)}$, we can get

$$s_{ij}^{LRW}=\frac{k_i}{2|E|}\cdot\pi_i(d_{ij})+\frac{k_j}{2|E|}\cdot\pi_j(d_{ji}) \quad (3)$$

Equation (3) represents the model which is called the local random walks based on the shortest path (LDRW) [12]. Using the shortest path in random walks is the most characteristic of this algorithm. At the same time, we introduce the concept of the first-time-passage probability based on the shortest path.

(2) KL divergence

Clustering is a group of points in Euclidean space R^n. The similarity of points within cluster is higher than the points outside cluster. In probability theory and information theory, the Kullback–Leibler divergence (KL divergence) is a non-symmetric measure of the difference between two probability distributions P and Q. P represents the distribution of observation data, and Q represents a theoretical model or a similarity distribution. The KL distance of the distribution Q from distribution$P$represents the extra needed information when the distribution Q replaces the distribution P. Although KL divergence usually regarded as a kind of distance, it is not a real distance, because it is not symmetrical. The KL distance of the distribution Q from distribution$P$is defined to be[13],

$$D_{KL}(P\|Q)=\sum_i P(i)\ln(\frac{P(i)}{Q(i)}) \quad (4)$$

3. Clustering Algorithm based on LDRW

(1) Construction of Stationary Markov chain

Using the idea of LDRW algorithm, we construct a complete graph G=(E,V) using the data set $X=\{x_n\}_{n=1}^{N}$. In the Markov chain, the probability from point $x_i$transferring to point x_j at steptis $P(x_j(t+1)|x_i(t))=P_{ij}$. The transition distribution of point$x_i$after steptis $P_i(t)=[p_{i1}(t),p_{i2}(t)...,p_{iN}(t)]$.

Definition 1: The subset C of the state space E is called closed set. Forarbitrary $i\in C$ and $j\notin C$, we can get $P_{ij}=0$. Obviously, the whole state space is a closed set. The closed set C is irreducible, if the state of the closed set C is interconnected.

Definition 2: In the homogeneous Markov chain, for arbitrary state $i\in E$ has a step set of random walks $\{n:p_{ii}^{(n)}>0,n\geq 1\}$, and the period of the state $i$is the greatest common divisor of this set. The state i is aperiodic if the greatest common divisor is 1. Because we consider about the similarity,the similarity is highest when the data point moves to itself. The walker must return to itself with a probability, so the steps set of random walks contains value 1. This can prove that the data point is aperiodic.

Definition 3: First-time-passage probability: The probability of the original state i reaching the state $j$after a state sequence for the first time is,

$$f_{ij}^{(n)}=P(T_{ij}=n\,|\,X_0=i)$$
$$=P(X_n=j,\ X_k\neq j,1\leq k\leq n-1\,|\,X_0=i).$$

While the probability of the state i finally reaching the state j after finite steps is,

$$f_{ij}=\sum_{n=1}^{\infty}f_{ij}^{(n)}.$$

The T_{ij} is the first time passage, namely the time or steps from one state to another for the first time in Markov chain. We suppose the original state is i and the final state is j, then the first time passage is $T_{ij}=\min\{n:X_0=i,\ X_n=j,n\geq 1\}$.

Definition 4: First-time-passage probability based on the shortest path: In the unweight graph, we determine the one-step transition probability matrix of the random walks according to the situation of link. If$p_{ab}$is nonzero then node$a$ andnode$b$ are linked. Supposed$_{ij}$ is the shortest path between state i and state j, then the $f_{ij}^{(n)}=0$ when n<d$_{ij}$. So the final arrival probability of two nodes is,

$$f_{ij}=\sum_{n=1}^{\infty}f_{ij}^{(n)}=\sum_{n=l_{ij}}^{\infty}f_{ij}^{(n)}.$$

Definition 5: Suppose the probability of state i returning to itself after step $n$is $f_{ii}=\sum_{n=1}^{\infty}f_{ii}^{(n)}$. The state i is recurrent if $f_{ii}=1$. Otherwise, the state i is non recurrent. In the state of recurrence, the average steps of walker returning to state i is represented by mathematical expectation $\mu_i=\sum_{n=1}^{\infty}nf_{ii}^{(n)}$. It is called positive recurrent if the state is recurrent and μ_i is a finite value. Otherwise, it is called null recurrent. The ergodic state is a positive recurrent and aperiodic state.

We can get the following two conclusions by the above definition.

1) The irreducible and finite Markov chain only has positive recurrent state.
2) The irreducible and aperiodic Markov chain is positive recurrent if and only if it has a stationary distribution.

Because the data points in the set of experimental data is finite, it forms a finite state set E. According to the Equation (4), if we regard all data points as a state space E, obviously, the state space E is a closed set and a complete graph, the Markov chain formed by data set is irreducible. According to the conclusion (1), the Markov chain only has positive recurrent state. And according to the conclusion (2), the Markov process will enter into a stable state $D_{KL}(P_i(t)\,|\,P_i(t+1))$ finally, namely the transition distribution $P_i(t)$ is convergent.

In this paper, the transition probability is a form of representation of similarity and the random walks is equivalently a map of metric space. Before random walking, all the data points and the similarity of points are in the same metric space. After random walking, the similarity distance between pointx_i and pointx_j will be changed. This change considers not only the original similarity of two points but also the similarity distance from other points to the pointx_j, and then compute the new similarity according to all these similarities. Because the time of each point into the steady state is different, we can use the transfer distribution of each point into the steady state be the final transfer distribution.

First, we compute the distribution list $P_i(1),P_i(2),\cdots$ of the point $x_i$according to the one-step transfer probability. Then using the KL distance to describe the changing process of distribution list when the point x_i in the process of random walks. The ratio of distribution distance, before and after one-step random walks, of the point x_i is$D_{KL}\big(P_i(t)|P_i(t+1)\big)$. We set a threshold value ε, when $D_{KL}\big(P_i(t)|P_i(t+1)\big)<\varepsilon$, the point x_i comes to steady state. So we can get the final distribution $P_i^{t_i}$ of each point.

(2) Clustering Analysis based on KL[15,17]

We assume that we can get K cluster $\{\mathrm{Q_k}\}_{\mathrm{k=1}}^{\mathrm{K}}$ after clustering data set X. The parameters $\mathrm{Q_k}$ is a centroid of a cluster. If there is a data point $\mathrm{P_i^{t_i}}\in\mathrm{Q_k}$, the amount in formation loss of $P_i^{t_i}$ can be represented as $D(P_i^{t_i}\,\|\,Q_k)$.

The objective function is the total loss of information when we use cluster centroid replaces the original data after clustering, denoted as $J(Q,I)=\sum_{k=1}^{K}\sum_{i\in I_k}D_{KL}(P_i^{t_i}\,\|\,Q_k)$. The goal of our algorithm is to ensure that the total loss of information is minimized. The method requires acontinuous calibration of each data partition.

First, we should give an initial centroid which can represent the original data well and the distance between centroids should as far as possible. We choose the even distribution of all data as the first centroid distribution. The other centroid distributions are determined by the maxmin rule.

Algorithm 1 Compute the initial centroid
Input: The data distribution $P_1^{t_1},P_2^{t_2},\cdots,P_n^{t_n}$ after being processed by the model of local random walks based on shortest path and the cluster number K.

Output: The initial centroid Q
Procedure:

(1) $Q_1=\dfrac{1}{N}\sum_{i=1}^{N}P_i^{t_i}$

(2)For k=2,3,…,K

$$z=\arg\max_i\min_{j=1,2,\ldots,k-1}D_{KL}(P_i^{t_i}\,\|\,Q_j),Q_k=P_z^{t_z}$$

And then is the clustering process. We partition the original data via the initial centroid, then adjust centroid of each cluster, and computation the centroid again. Repeat the procedure until the objective function $J(Q,I)$ does not decrease.

Algorithm 2:
Input: The data distribution $P_1^{t_1},P_2^{t_2},\cdots,P_n^{t_n}$ after being processed by the model of local random walks based on shortest path and the cluster number K.

Output: The final result of clustering and the centroid distribution Q

Procedure:

(1) For each data point, compute the similarity of each to the centroid and merge it into the most similar clusters. Then we get a new partition $I_k^{'}=\{i:k=\arg\min_k D_{kl}(p_i^{t_i}\,\|\,Q_k)\}$.

(2) Update each centroid of cluster. For $k=1,2,\ldots,K$,

$$Q^{'}=\frac{1}{|I_k^{'}|}\sum_{i\in I_k^{'}}p_i^{t_i}$$

(3) If $J(Q,I)\geq J(Q^{'},I^{'})$, $Q=Q^{'}$, then go to the first step.

Otherwise, loop end.

In the above algorithm, we transfer the metric space of data point. So we make the distance of each two point reference other distance. We finally determine each distribution through the random walks, and utilize the total loss of information to be the objective function.

4. Result and Analysis

(1) The experimental data and processing

We use the UCI handwritten numeral lattice diagram data to verify our algorithm [14]. Every data in the data set is a 32 32two-value matrix contained handwritten numeral which is from 0 to 9. Every matrix contains a correct label.

Suppose $X=\{x_n\}_{n=1}^{N}$ is the experiment data set, and N is the number of the data set. First we define similarity of two-value matrix as the proportion of the same number of elements in the two matrices of the same position. If we use set difference to define the similarity between points, points in the data set are some sets which contain 32×32 elements which are different between each other according to their positions. So the definition of the similarity is,

$$d(x_i,x_j)=1-\frac{\|x_i-x_j\|}{\|x\|}$$

Because any number of a data element is a 32×32 matrix, we define an uniform value$\|x\|$. We can get the similarity matrix$D(d_{ij})$by computing the similarity of all data point pairs. We define the transition probability of from point x_i to point x_j as $p_{ij}=d_{ij}\Big/\sum_{i=1}^{N}d_{ij}$. We can find that the higher the similarity between two points, the higher the transition probability between them.

(2) Analysis of Experimental Results

For verifying the property of clustering algorithm proposed in this paper, the K-means algorithm which is used to be comparison do the same step with the original K-means algorithm and the improved K-means algorithm except finding final distribution without random walks. Every original data point contains a 32×32 handwritten numeral lattice diagram and a label. This label represents the real number of this graph. After clustering we get 10 clusters, and confirm their label according to the majority principle. For example, if the majority of the original labels of a cluster are 0, the label of this cluster is 0. Meanwhile the label of all the experiment data points in the cluster is 0. If the final classify of a data point is same as its original label, it is a correct classification. Counting all the data points which are correct classifications and comparing with all the data points can get the accuracy. The result indicates that the accuracy of using K-means algorithm directly is 73%, but the accuracy can be raised to 75% when we use random walks to determine the stable distribution of data point. As shown in Fig.1.

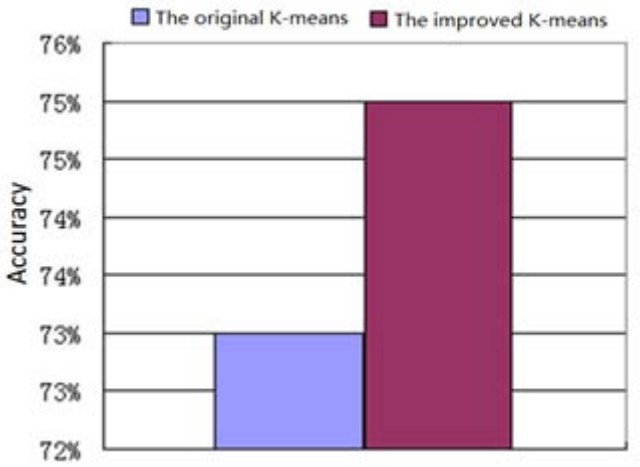

Figure 1: Comparison of the two algorithm

The general K-means algorithms are just need to save data points and centroids. If n is the number of property, the storage space is $O((m+K)n)$. If we suppose the number of clustering iterations is$I$, the time complexity is $O(I\times K\times m\times n)$. The clustering process of most data points have been completed in the first few times. If the threshold value is chosen well, we can get a balance between time complexity and the effect of algorithm. For the new K-means algorithm, we should add storage $O(m^2)$ for a one-step transition probability matrix. And the average time complexity degree of random walks is the average shortest path.

5. Conclusion

In this paper, we propose a clustering algorithm based on link data and random walks. It uses the directly linking between data points to form a network, and refers the link information between data point and other data points when we define the similarity. Then we use k-means algorithm to perform test. Experimental results show that it is benefit to find the real similarity of data points when we consider the network characteristic of data sufficiently.

6. Acknowledgements

This research was supported by the National Natural Science Foundation of China (NSFC61572005, NSFC61370060), and the Fundamental Research Funds for the Central Universities (2016JBM019,2015JBM035).

References

[1] Szekely, Gabor J., and Maria L. Rizzo. "Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method."Journal of classification22, no. 2 (2005): 151-183.
[2] Fernández, Alberto, and Sergio Gómez. "Solving non-uniqueness in agglomerative hierarchical

clustering using multidendrograms."Journal of Classification25, no. 1 (2008): 43-65.
[3] Gao, Hui, Jun Jiang, Li She, and Yan Fu. "A New Agglomerative Hierarchical Clustering Algorithm Implementation based on the Map Reduce Framework."JDCTA4, no. 3 (2010): 95-100.
[4] Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm."Applied statistics(1979): 100-108.
[5] Taskar, Benjamin, Eran Segal, and Daphne Koller. "Probabilistic classification and clustering in relational data." InInternational Joint Conference on Artificial Intelligence, vol. 17, no. 1, pp. 870-878. LAWRENCE ERLBAUM ASSOCIATES LTD, 2001.
[6] Taskar, Ben, Pieter Abbeel, and Daphne Koller. "Discriminative probabilistic models for relational data." InProceedingsoftheEighteenthconferenceonUncertaintyinartificialintelligence,pp.485-492.MorganKaufmann PublishersInc.,2002.
[7] Holder,LawrenceB.,andDianeJ.Cook."Graph-BasedDataMining."Encyclopediaofdatawarehousingandmining 2(2009):943-949.
[8] Caldarelli,Guido,andAlessandroVespignani.(editedby) Largescalestructureanddynamicsofcomplexnetworks:frominformationtechnologytofinanceandnaturalscience. Vol.2.WorldScientific,2007.
[9] MacQueen,James."Somemethodsforclassificationandanalysisofmultivariateobservations."InProceedingsofthefifthBerkeleysymposiumonmathematicalstatisticsandprobability,vol.1,no.14,pp.281-297.1967.
[10] Steinhaus,H."Surladivisiondescorpsmatérielsenparties "(inFrench).Bull.Acad.Polon.Sci.4(12):801-804.1957.
[11] Lü,Linyuan,andTaoZhou."Linkpredictioninweightednetworks:Theroleofweakties."EPL(EurophysicsLetters) 89,no.1(2010):18001.
[12] Xu,Baomin,TinglinXin,YunfengWang,andYanpinZhao."LocalRandomWalkwithDistanceMeasure."Modern PhysicsLettersB27,no.08(2013).
[13] http://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence
[14] http://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits
[15] HEHuimin.ClusteringAlgorithmBasedonRandomWalk ModelandKL-divergence.ComputerEngineering.34(16):224-226.2008.
[16] Harel,David,andYehudaKoren."Onclusteringusingrandomwalks."InFSTTCS2001:FoundationsofSoftwareTechnologyandTheoreticalComputerScience,pp.18-41.SpringerBerlinHeidelberg,2001.
[17] Maranzana,F.E."Onthelocationofsupplypointstominimizetransportcosts."OR(1964):261-270.