Density Based Clustering Using Gaussian Estimation Technique

R.Prabahari[1], Dr.V.Thiagarasu[2]
Assistant Professor[1], Associate Professor[2]
PG & Research Department of Computer Science,
Gobi Arts & Science College,
Gobi – 638 452, Tamil Nadu, India.
Email:senthil.praba11@gmaill.com, profdravt@gmail.com

***Abstract:* -** Density based clustering algorithm (DENCLUE) is one of the primary methods for clustering in data mining. The clusters which are formed based on the density are easy to understand and it does not limit itself to the shapes of clusters. The Denclue algorithm employs a cluster model based on kernel density estimation and a cluster is densed by a local maximum of the estimated density function. Data points are assigned to clusters by hill climbing, i.e. points going to the same local maximum are put into the same cluster. The traditional density estimation is only consider the location of the point, not variable of interest and hill climbing makes unnecessary small steps in the beginning and never converges exactly to the maximum. This paper proposes an improved hill climbing method. The density is measured using influence function and variable interest. The proposed method forms cluster by associating points with density attractors.
The clusters are formed well-defined. So the outliers can be detected in efficiently.

Keywords: *Clustering, Density based clustering, DENCLUE, OPTICS, DBSCAN*

*****

1. Introduction

Density-based clustering methods group neighboring objects into clusters based on local density conditions rather than proximity between objects [Deepti Sisodia et al, 2012]. These methods regard clusters as dense regions being separated by low density noisy regions. Clustering can be formulated in many different ways. .Non-parametric methods are well suited for exploring clusters, because no generative model of the data is assumed. Instead, the probability density in the data space is directly estimated from data instances. Kernel density estimation [Parimala M et al., 2011] is a principled way of doing that task. There are several clustering algorithms, which exploit the adaptive nature of a kernel density estimate. However, the algorithms use directional information of the gradient only. The step size remains fixed throughout the hill climbing. This implies certain disadvantages, namely the hill climbing does not converges towards the local maximum, it just comes close, and the number of iteration steps may be large due to many unnecessary small steps in the beginning. The step size could be heuristically adjusted by probing the density function at several positions in the direction of the gradient. As the computation of the density function is relatively costly, such a method involves extra costs for step size adjustment, which are not guaranteed to be compensated by less iteration.

The contribution of this article is an improved hill climbing method for kernel density estimates with Gaussian kernels. The method adjusts the step size automatically at no additional costs and converges towards a local maximum. This paper proves the hill climbing as a special case of the expectation maximization algorithm. Depending on the convergence criteria, the method needs less iterations as fixed step size methods. Variants of density based clustering are DBSCAN [Derya Birant, Alp Kut.2010], OPTICS [Ankerst M et al, 1999], and follow up versions, which do not use a probabilistic framework.

This paper is organized as follows. Literature surveys are given in section 2. In section 3 the proposed algorithm is designed to discover the density based clusters. Experimental results are reported in section 4. Conclusions are presented in section 5.

2. Literature Survey

Density-based clustering method regard clusters as dense regions being separated by low density noisy regions. Density-based methods have noise tolerance, and can discover non-convex clusters. Similar to hierarchical and partitioning methods, density-based techniques encounter difficulties in high dimensional spaces because of the inherent scarcity of the feature space, which in turn, reduces any clustering tendency [Anant Ram et al.,2010]. The representative examples of density based clustering algorithms are [Sander J et al., 1998] DBSCAN, OPTICS and DENCLUE.

2.1 DENCLUE

Density-based Clustering (DENCLUE) uses an influence function to describe the impact of a point about its neighborhood while the overall density of the data space is the sum of influence functions from all data. Clusters are determined using density attractors, local maxima of the overall density function. To compute the sum of influence functions a grid structure is used. DENCLUE scales well, can find arbitrary-shaped clusters, is noise resistant, is insensitive to the data ordering, but suffers from its sensitivity to the input parameters. The dimensionality phenomenon heavily affects Denclue's effectiveness [Chen Ning et al., 2002]. Moreover, similar to hierarchical and partitioning techniques, the output of density-based methods cannot be easily assimilated by humans. In short, the advantages of density-based clustering are discovery of arbitrary-shaped clusters with varying size and resistance to noise and outliers. The disadvantages of density-based clustering are high sensitivity to the setting of input parameters, poor cluster descriptors and Unsuitable for high-dimensional datasets because of the dimensionality phenomenon.

2.1.1 DENCLUE ALGORITHM

Step 1: Find the influence of each data point can be modeled using a Gaussian influence function

Step 2: The overall density of the data space can be modeled analytically as the sum of the influence function applied to all data points

Step 3: Clusters can then be determined by identifying density attractors where density attractors are local maximum of the overall density function.

2.1.2 DENSITY ATTRACTOR

A clustering in the Denclue 1.0 framework is defined by the local maxima of the estimated density function. A hill-climbing procedure is started for each data instance, which assigns the instance to local maxima. In case of Gaussian kernels, the hill climbing is guided by the gradient of $\hat{p}(x)$. The hill climbing procedure starts at a data point and iterates until the density does not grow anymore. The step size δ is a small positive number. In the end, those end points of the hill climbing iteration, which are closer than 2δ, are considered, to belong to the same local maximum [Mumtaz K and K. Duraiswamy,2012] Instances, which are assigned to the same local maximum, are put into the same cluster. A practical problem of gradient based hill climbing in general is the adaptation of the step size i.e. how far to follow the direction of the gradient. There are several general heuristics for this problem, which all need to calculate $\hat{p}(x)$ several times to decide a suitable step size. In the presence of random noise in the data, the Denclue framework provides an extra parameter ξ > 0, which treats all points assigned to local maxima with $p(x) < \xi$ as outliers.

3 PROPOSD SYSTEMS: DENCLUE

In this section, propose significant improvements of the traditional algorithm for Gaussian kernels. Since the choice of the kernel type does not have large effects on the results in the typical case, the restriction on Gaussian kernels is not very serious. First, introduce a hill climbing procedure for Gaussian kernels, which adjust the step size automatically at no extra costs [Joanna Tan and Jing Zhang 2010]. The method does really converge towards a local maximum.

Throughout the paper, assume that datasets have the form <Location>, <variable_of_interest>). More formally, a dataset O is a set of data objects, where n is the number of objects in O belonging to a feature space F.

$$O = \{o_1, o_2, o_3, \ldots, o_n\} \,\varepsilon\, F \qquad (3.1)$$

assume that objects $o \,\varepsilon\, O$ have the form $((x, y), z)$ where (x, y) is the location of object o, and z - denoted as $z(o)$ is the value of the variable of interest of object o. The variable of interest can be continuous or categorical. Besides, the distance between two objects in O, $o_1=((x_1, y_1), z_1)$ and $o_2=((x_2, y_2), z_2)$ is measured as $d((x_1, y_1), x_2, y_2))$ where d denotes a Euclidian distance. In the following, this paper will introduce density estimation techniques. Density estimation is further subdivided into categorical density estimation, and continuous density estimation,[He Zengyou et al., 2002] depending whether the variable z is categorical or continuous.

3.1 Influence and Density Functions

In general, density estimation techniques employ influence functions that measure the influence of a point $o \,\varepsilon\, O$ with respect to another point $v \,\varepsilon\, F$, a point o's influence on another point v's density decreases as the distance between o and v increases. In particular, the influence of object o ε O on a point $v \,\varepsilon\, F$ is defined as:

$$f_{\text{influnece}}(v,o) = z(o) * e \frac{-d(v,o)^2}{2\sigma^2} \qquad (3.2)$$

If for every $o \,\varepsilon\, O$ $z(o)=1$ holds, the above influence function become a Gaussian kernel function, commonly used for density estimation and by the density-based clustering algorithm DENCLUE[Murray et al., 1998]. The parameter σ determines how quickly the influence of o on v decreases as the distance between o and v increases.

The overall influence of all data objects $o \,\varepsilon\, O$ on a point $v \,\varepsilon\, F$ is measured by the density function $f^o(v)$, which is defined as follows:

$$f^o(v) = \sum f_{influnece}(v,o) \qquad (3.3)$$

In categorical density estimation[He Zengyou et al.,2002], assume that the variable of interest is categorical and takes just two values that are determined by the membership in a class of interest[Mumtaz K et al.,2012]. In this case, $z(o)$ is defined as follows:

$$z(o) = \begin{cases} 1 \text{ if o belong to the class of interest} \\ -1 \text{ otherwise} \end{cases} \qquad (3.4)$$

Here the objects in data set belong to any one of two classes.

3.2 Local maximum procedure

The proposed algorithm operates on the top of the influence and density functions that were introduced in formulas [Prabahari R and Thiagarasu V, 2014] and [Martin Ester et al.,]. Its clustering process is a hill-climbing process that computes density attractors. During the density attractor calculation process, data objects are associated with density attractors forming clusters.

A point, a, is called a density attractor of a dataset O if and only if it is a local maximum of the density function f^o and $| f^o(a)| > \xi$□, where ξ is a density threshold parameter. For any continuous and differentiable influence function, the density attractors can be calculated by using hill-climbing procedure.

The proposed algorithm does not calculate the density attractor for every data object. During the density attractor calculation, it examines data objects close to the current point when moving towards the supervised density attractor. If their density values have the same sign as the current density value at the calculating point, these data objects will be associated with the same final density attractors after it has been computed.

. In each iteration step, locate the data objects close to o_i^{j+1}, i.e. data objects whose distance to o_i^{j+1} is less than σ/2. Each data object close to o_i^{j+1} is processed as follows:

If the data object does not belong to any cluster yet, the object is marked with the same cluster ID as those attracted by the current attractor computation.

If the data object already belongs to a cluster $c_{i,}$ all the data points in the cluster c_i are marked with the same cluster ID as those objects that have been collected by the current attractor computation.

The hill climbing procedure stops returning a density attractor a; if $|f^{O}(a)| > \xi$, a is considered a density attractor, and a cluster is formed.

The proposed algorithm uses the following formula to find the direction of the object.

$$dx=f(x+step,y)-f(x,y)$$

$$dy=f(x,y+step)-f\phi(x,y)$$

$$\nabla f(x,y)=\nabla f(u)=(dx/(\|(dx,dy)\|),dy/(|dx|+|dy|) \qquad (3.5)$$

It is important to stress that DENCLUE's hill climbing procedure uses the influence function and not the density function to determine in which direction to proceed.

$$\nabla\Psi(u)=\sum_{v\in neighbor(u)}(v-u)*f_{influence}(u,v)$$

(3.6)

However, this formula should be normalized–otherwise, far away points would receive a higher weight in direction computations–obtaining formula. This approach considers all the points in the neighborhood of an iteration point, and the influence of a point in the neighborhood decreases as the distance to the iteration point increases.

$$\nabla\Psi(u)=\sum_{v\in neighbor(u)}(v-u)*f_{influnece}(u,v)/\|v-u\| \qquad (3.7)$$

The most critical parameter when using the proposed is σ, which determines the size of the influence region of an object. Larger values for σ increases the influence regions and more neighboring objects are affected. From a global point of view, larger values for σ results more objects are connected and joined together by density paths between them. Therefore when σ decreases, there are more local maxima and minima and the number of clusters increases. To select proper parameters, to visualize the density function and/or create a density contour map of the dataset for some initial values for σ and ξ first, and then use the displays to guide the search for good parameter values. Sometimes, the domain experts give the approximate number of clusters or number of outliers; this information can be used to determine good values for σ and ξ automatically using a simple binary search algorithm. Based on past experience, choosing good values for the other parameters of the proposed is straight forward.

3.3. Run Time Complexity

The run time of the proposed method depends on algorithm parameters, such as σ,ξ, ω and step. The grid partitioning part takes *O(n)* in the worst case, where n is the total number of data objects in the dataset. The run time for calculating the density attractors and clustering varies for different values of parameters σ and ξ. It will not go beyond *O(h*n)* where *h* is the average hill-climbing time for a data object. The total run time is approximately *O(n + h*n).*

4 Result and Discussions

In proposed system the location and variable of interest is considered for clustering. Also, the randomly created data set is also used to test and validate the proposed system. The table shows that run time of proposed system is lowest while other denclue having highest run time. In terms of cluster quality, the proposed algorithm leads while other versions of denclue lacking behind.When using categorical density estimation techniques, decision boundaries between the two classes represent areas whose density is equal to 0.

Algorithms	Input para meters	Noise Handle	Run time (ms)	Cluster Quality
Denclue1	Two	Good	40	91.3%
Denclue2	Two	Good	38	92.5%
Proposed	Two	Very Good	30	97.08%

The key role of proposed algorithm is to find the density attractor which is shown in Figure 4.1.

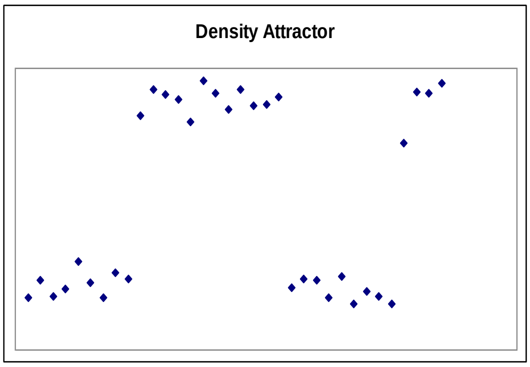

Fig 4.1: DENCLUE Density Attractor

Conclusions

This paper proposes a density estimation approach that extends the traditional density estimation techniques by considering a variable of interest that is associated with a spatial object. Density is measured as the product of an influence function with the variable of interest. The proposed algorithm uses local maximum method to calculate the maximum/minimum (density attractors) of a density function and clusters are formed by associating data objects with density attractors during the local maximum procedure. Compared with other algorithms the proposed method performed quite well.

References

[1] Anant Ram, Sunita Jalal, Anand S. Jalal, Manoj kumar, "A density Based Algorithm for Discovery Density Varied cluster in Large spatial Databases*",* International Journal of Computer Application Vol. 3,No.6, June 2010.

[2] Ankerst M, M. M. Breunig, H.P. Kriegel, and J. Sander, "Optics: Ordering points to identify the clustering structure", In *Proceedings* SIGMOD '99, pp 49-60, ACM Press, 1999.

[3] Chen Ning, Chen An, Zhou Longxiang, "An Incremental grid Density Based Clustering Algorithm", Journal of Software, 13(1), pp. 1-7, 2002.

[4] Deepti Sisodia, Lokesh Singh, Sheetal Sisodia, Khushboo saxena, "Clustering Techniques: A Brief Survey of Different Clustering Algorithms", International Journal of Latest Trends in Engineering and Technology, Vol. 1 Issue 3 pp. 82-87 September 2012

[5] Derya Birant, Alp Kut, "ST-DBSCAN: An Algorithm for Clustering Spatial-temporal data" Data and Knowledge Engineering, pp. 208-221, 2007.

[6] He Zengyou, Xu Xiaofei , Deng Shengchun, Squeezer, "An efficient algorithm for clustering categorical data", Journal of Computer Science and Technology, pp. 611-624, May 2002.

[7] Jianhao Tan and Jing Zhang "An Improved Clustering Algorithm Based on Density Distribution Function" Computer and Information Science Vol. 3, No. 3; August 2010

[8] Martin Ester,Han-peter Kriegel,Jorg Sander, Xiaowei Xu,"A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", *2nd International conference on Knowledge Discovery and Data Mining (KDD-96)*

[9] Mumtaz K, Dr. K. Duraiswamy, "An Analysis on Density Based Clustering of Multidimensional Spatial Data", Indian Journal of Computer Science and Engineering Vol. 1 pp. 8-12

[10] Murray, A. T. and Estivill-Castro, V. Cluster discovery techniques for exploratory spatial data analysis, International Journal of Geographical Information Science, Vol. 12, No. 5, 1998, 431-443.

[11] Parimala M, D. Lopez, N. C. Senthilkumar, "A Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases", International Journal of Advanced Science and Technology, Vol. 31, June 2011.

[12] Prabahari R, Dr.V.Thiagarasu, "A Comparative Analysis of Density Based Clustering Techniques for Outlier Mining", International Journal Of Engineering Sciences & Research Technology, ISSN 2277-9655, pp 132-136 November 2014.

[13] Sander J, M. Ester, H. P. Kriegel, and X. Xu, "Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications", Data Mining and Knowledge Discovery, pp 169-194, 1997.

[14] Sander, J., Ester, M., Kriegel, H.P., and Xu, X., Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications. *Data Mining and Knowledge Discovery*, Kluwer Academic Publishers, Vol. 2, No. 2, 1998, 169-194

[15] Zhou Yonggeng, Zhou Aoying, Cao Jing, "DBSCAN Algorithm Based Data Partition", Journal of Computer Research and Development, 37(10), pp 1153-1159 , 2000.