BEST: International Journal of Management, Information
Technology and Engineering (BEST: IJMITE)
ISSN 2348-0513
Vol. 2, Issue 6, Jun 2014, 21-26
© BEST Journals

ACTIVITIY PREDICTION FOR DIFFERENTLY-ABLED PEOPLE USING DBSCAN AND GAUSSIAN DISTRIBUTION

JANHAVI S[1], BHAGYASHREE S[2], ANJANA AV[3] & MOUNESHACHARI S[4]

[1 2 3]Final Year Student, Bachelor of Engineering, Department of Computer Science & Engnieering,
GSSSIETW, Mysore, Karnatka, India

[4]Associate Professor, Department of, Computer Science & Engnieering GSSSIETW, Mysore, Karnataka, India

ABSTRACT

The main objective of this paper is to train systems such that the systems itself will guide the differently abled people to perform their activities(tea, coffee, lunch & etc…). This paper provides a way in which system will remind the person to perform the activity which is adaptive to the behavior changes by combining classification and clustering techniques. By using DBSCAN algorithm the data is clustered. DBSCAN is a density based spatial clustering of applications with noise in which it finds a number of clusters starting from the estimated density distribution of corresponding nodes. After grouping the data, the Gaussian distribution algorithm is used to find the routine performed by the differently-abled people at a maximum extent. Gaussian distribution is a very commonly occurring continuous probability distribution —a function that tells the probability that any real observation will fall between any two real limits or real numbers as the curve approaches zero on either side.

KEYWORDS: Differently-Abled People, DBSCAN, Gaussian Distribution, Clustering

INTRODUCTION

The problems faced by the "Differently-able People" are social interaction, language and communication. We present this paper which will minimize the problem to some extent. Due to busy schedule of human life, people are forgetting their daily routine. This paper will bring a solution for it in which system itself intimates the daily routine of the differently-abled people. The daily routine of the differently-abled people is been observed and stored in the logfile. The logfile consists of activity performed, time at which the activity is performed. Conventional Database methods are inadequate to extract. useful information from huge data banks. Cluster analysis is one of the major data analysis methods and the DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature.

REVIEW OF LITERATURE

Clustering Algorithm

Cluster analysis is the organization of a set of objects into classes or clusters based on similarity. Intuitively, objects within a valid cluster are more similar to each other than they are to an object belonging to a different cluster. The variety of techniques for representing data, measuring proximity (similarity) between data elements, and grouping data elements have produced a rich and often confusing assortment of clustering methods. It is important to understand the difference between clustering (unsupervised classification) and discriminate analysis (supervised classification) [1]. Many researchers have defined four steps for cluster analysis: feature selection or extraction, cluster algorithm design and selection, cluster validation, and result interpretation. These steps are closely related to each other and affect the derived clusters. Several researchers have given significant contribution on the study of cluster techniques.

Roughly, these clustering algorithms can be separated into five general categories [1]; hierarchical clustering, partition clustering, grid-based clustering, model-based clustering, and density-based clustering.

Density Based Clustering

Density based clustering methods discover cluster based on the density of points in regions. Therefore density based clustering methods are capable to produce arbitrary shapes clusters and filter out noise (outlier) [2] [3]. Ester et al [2] introduced density based algorithms DBSCAN and further it was generalized [3] by using symmetric and reflexive binary predicate and introduces some non-spatial parameter ―cardinality‖. Thus the GDBSCAN algorithm can cluster point objects as well as spatially extended objects according to both, their spatial and their non-spatial, attributes. Apart from this, several variants of DBSCAN algorithm have been reported in literature. The key feature of DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is that for each object of a cluster, the neighborhood of a given radius ε has to contain at least a specified minimum number MinC of objects, i.e., the cardinality of the neighborhood has to exceed a given threshold. Radius ε and minimum number MinC of objects are specified by user. Let D be a data set of objects, the distance function between the objects of D is denoted by DIST and given parameters are ε and MinC then DBSCAN can be specified by the following definitions. The following definitions are adopted from Ester et al [2].

Definition 1: (Neighbourhood of an object). The εneighbourhood of an object p, denoted by Nε (p) is defined by N ε (p) = {q ϵ D | DIST (p, q) <= ε}.

Definition 2: (Direct Density Reachability). An object p is direct density reachable from object q w.r.t. ε and MinC if |Nε(P)| >= MinC ʌ p ϵ Nε (q). q is called core object ε when the condition |Nε (P)| >= MinC holds.

Definition 3: (Density Reachability). An object p is densityreachable from an object q w.r.t ε and MinC if there is a sequence of objects p1…pn; p1 = q, pn = p such that pi+1 is direct density reachable from pi.

DBSCAN chooses an arbitrary object p. It begins by performing a region query, which finds the neighborhood of point q. If the neighborhood contains less than MinC objects, then object p is classified as noise. Otherwise, a cluster is created and all objects in p's neighborhood are placed in this cluster. Then the neighborhood of each of p's neighbors is examined to see if it can be added to the cluster. If so, the process is repeated for every point in this neighborhood, and so on. If a cluster cannot be expanded further, DBSCAN chooses another arbitrary unclassified object and repeats the same process. This procedure is iterated until all objects in the dataset have been placed in clusters or classified as noise.

Gaussian Distribution Algorithm

In probability theory, the normal (or Gaussian) distribution is a very commonly occurring continuous probability distribution—a function that tells the probability that any real observation will fall between any two real limits or real numbers, as the curve approaches zero on either side. Normal distributions are extremely important in statistics and are often used in the natural and social sciences for real-valued random variables whose distributions are not known. [4]

The normal distribution is immensely useful because of the central limit theorem, which states that, under mild conditions, the mean of many random variables independently drawn from the same distribution is distributed approximately normally, irrespective of the form of the original distribution: physical quantities that are expected to be the sum of many independent processes (such as measurement errors) often have a distribution very close to the normal. Moreover, many results and methods (such as propagation of uncertainty and least squares parameter fitting) can be derived analytically in explicit form when the relevant variables are normally distributed.

The Gaussian distribution is sometimes informally called the bell curve as shown in Fig 1. However, many other distributions are bell-shaped (such as Cauchy's, Student's, and logistic). The terms Gaussian function and Gaussian bell curve are also ambiguous because they sometimes refer to multiples of the normal distribution that cannot be directly interpreted in terms of probabilities. A normal distribution is

$$f(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}e^{\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

The parameter μ in this definition is the mean or expectation of the distribution (and also its median and mode). The parameter σ is its standard deviation; its variance is therefore σ2. A random variable with a Gaussian distribution is said to be normally distributed and is called a normal deviate.

$$\text{Z-Score} = (x- \mu)/ \sigma \tag{2}$$

If μ = 0 and σ = 1, the distribution is called the standard normal distribution or the unit normal distribution, and a random variable with that distribution is a standard normal deviate.

The normal distribution is the only absolutely continuous distribution all of whose cumulates beyond the first two (i.e., other than the mean and variance) are zero. It is also the continuous distribution with the maximum entropy for a given mean and variance. [5][6]

The Gaussian distribution belongs to the family of stable distributions which are the attractors of sums of independent, identically distributed distributions whether or not the mean or variance is finite. Except for the Gaussian which is a limiting case, all stable distributions have heavy tails and infinite variance.

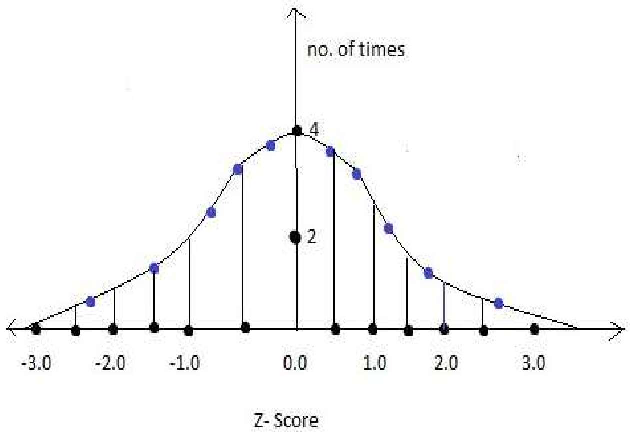

Figure 1: Gaussian Bell Curve

EXPERIMENTAL RESUTLS

Data Pre-Processing

For the task of clustering we are using the data stored in the log file of a system as shown in table 1. Data consists of the activity performed by the differently-abled people and the time at which the activity was performed between April 1, 2014 to April 7, 2014. Data have been recorded only when system was used. Each record consists of the activity performed and the time at which the activity was performed. For the recorded data the DBSCAN algorithm is applied to form 8 clusters as shown in Figure 2.

Table 1: Sample Log File

Time(min)	Activities	Time(min)	Activities
20	A1	100	A1
30	A1	101	A2
50	A1	110	A2
70	A2	120	A2
80	A2	130	A1
90	A1	150	A1
91	A2	160	A2

Time(min)	Activities	Time(min)	Activities
200	A3	170	A2
210	A1	180	A1
220	A1	181	A2
221	A3	182	A3
230	A2	190	A1
235	A3	191	A2
240	A2	192	A3

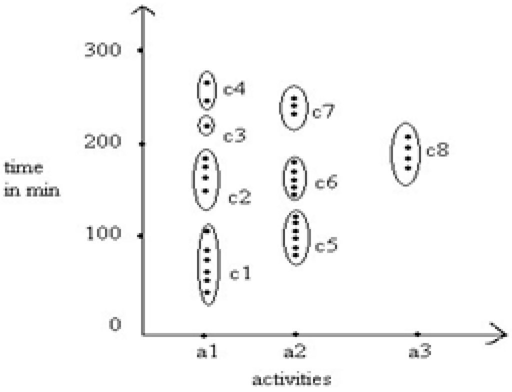

Figure 2: Clusters of Activities

For each cluster Gaussian distribution algorithm is applied to find the time at which the activity is performed maximum number of times. So that the system itself can intimate the differently-abled people to perform the particular activity at that time. In our paper we have shown the Gaussian distribution of few clusters. Figure 3, 4, 5 and 6 shows the Gaussian distribution of clusters c1, c2, c6, c8 respectively. Similarly the Gaussian distribution can be applied for remaining clusters.

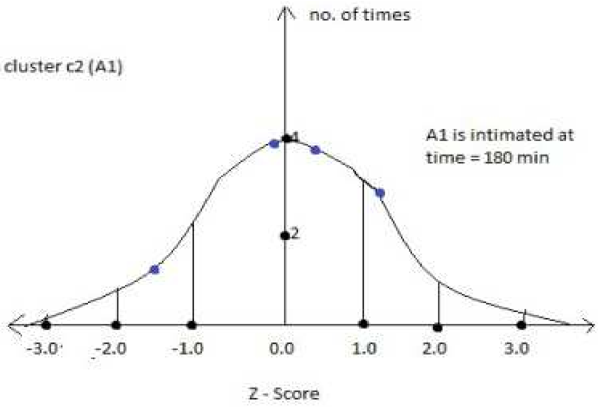

Figure 3: Bell Curve of Cluster c1

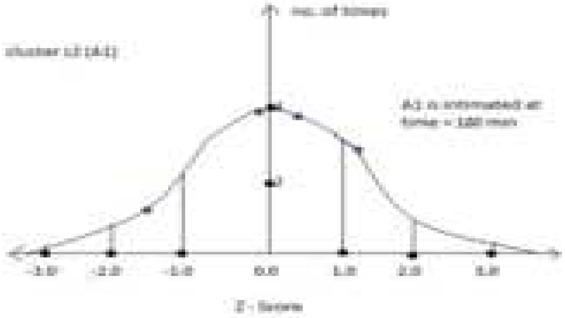

Figure 4: Bell Curve of Cluster c2

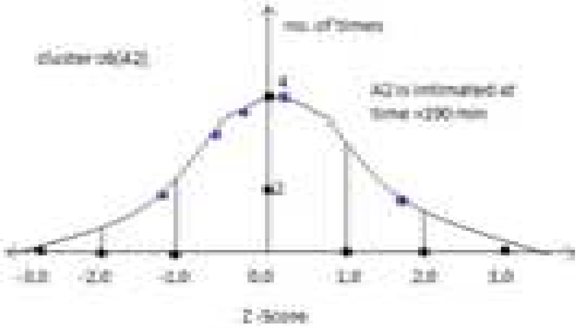

Figure 5: Bell Curve of Cluster c6

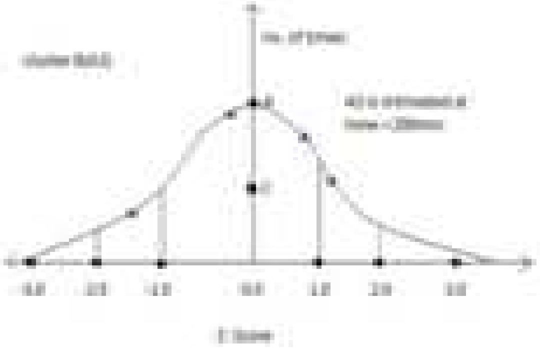

Figure 6: Bell Curve of Cluster c8

From the bell curves of the clusters above, the experimental result is shown in the below table 2.

Table 2: Experimental Results

Clusters	Experimental Result
Cluster c1	Intimates activity A1 at time = 50 min
Cluster c2	Intimates activity A1 at time = 180 min
Cluster c6	Intimates activity A2 at time = 190 min
Cluster c8	Intimates activity A3 at time = 200 min

Table 3

Time (min)	Z-Score
20	-1.26
30	-1.01
50	-0.5
90	0.75
100	0.76
130	1.51

Table 4

Time (min)	Z –Score
160	-1.07
170	-0.66
180	-0.24
190	0.165
230	1.821

Table 5

Time (min)	Z- Score
150	-1.5
180	-0.1
190	0.34
210	1.27

Table 6

Time (min)	Z-Score
180	-1.43
200	-0.39
220	0.65
230	1.17

CONCLUSIONS

In this paper we demonstrated that how density based clustering and Gaussian distribution can be used to guide the differently-abled people to perform their daily routine. We believe that it will bring a revolutionary change so that the differently-abled people can perform their activity independently. In our future work we will try to fill the communication gap between the differently-abled people and others.

REFERENCES

1. R. Xu, and C. D. Wunsch, C. D., ‖Survey of Clustering Algorithm‖, IEEE Trans. of Neural Networks, Vol. 16, No. 3, pp. 645-678, 2005.
2. M. Ester, H. P. Kriegel, J. Sander and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", In: Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD‟96), Portland, AAAI Press pp. 291-316, 1996.
3. J. Sander, M. Ester, H. P. Kriegeland X. Xu, "Density- Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications", Int. J. of Data Mining and Knowledge Discovery, Kluwer Academic Publishers vol. 2, pp. 169-194, 1998.
4. Normal Distribution, Gale Encyclopedia of Psychology, Casella & Berger (2001, p. 102)
5. Cover, Thomas M.; Thomas, Joy A. (2006). Elements of Information Theory. John Wiley and Sons. p. 254.
6. Park, Sung Y.; Bera, Anil K. (2009). "Maximum Entropy Autoregressive
7. Conditional Heteroskedasticity Model". Journal of Econometrics (Elsevier) 150(2): 219–230. doi:10.1016/j.jeconom.2008.12.014. Retrieved 2011-06-02. Study of Michael Okpara University of Agriculture, Umudike. Library Philosophy and Practice 1-8.