A survey paper on k mean clustering cultural studies essay

Pritesh Vora#, Bhavesh Oza*,#PG Student, Information technology Department, L. D. College of Engineering, Ahmedabad (GTU)*Assistant Prof., Computer Engineering Department, L. D. College of Engineering, Ahmedabad (GTU)1Pritesh2212@gmail. com2bhavesh_oza_2001@yahoo. co. inAbstract— In Data Mining Clustering is an important research topic and wide range of unsupervised classification application. Clustering is technique which divide a data into meaningful groups. K-mean one of the popular clustering algorithm. K-mean clustering widely used to minimize the squared distance between points in the same cluster. Particle swarm optimization is an evolutionary computation techniques which find a optimum solution in many application. Using the PSO optimized clustering results in the components, in order to get a more precise clustering efficiency. In this paper present the comparison of K-mean clustering and the Particle swarm optimization.Keywords— Clustering, K-mean Clustering, Particle Swarm OptimizationIntroductionClustering is a technique that can divide data objects in to groups based on information found in the data that describes the objects and their relationships, which can be used in many applications, such as data mining and knowledge discovery, vector quantization, pattern recognition, and etc. There are two main techniques in clustering known as hierarchical and partitional clustering. In hierarchical clustering, the data are not partitioned into a particular cluster in a single step, instead, a series of partitions takes place, which may run from a single cluster containing all objects to n clusters each containing a single object. And each cluster can have sub clusters, so it can be viewed as a tree, a node in the tree is a cluster, the root of the tree is the cluster containing all the objects, and each node, except the leaf nodes, is the union of its children. But in partitional

clustering, the algorithms typically determine all clusters at once, it divides the set of data objects into non-overlapping clusters, and each data object is in exactly one cluster. The partitional clustering can be used as divisive algorithms in the hierarchical clustering. Particle swarm optimization (PSO) has gained much attention, and it has been applied in many fields. PSO is a useful stochastic optimization algorithm based on population . The birds in a flock are represented as particles, and particles can be considered as simple agents flying through a problem space. And the particle’s location in the multi-dimensional problem space can represent the solution for the problem. But the PSO may lack global search ability at the end of a run due to the utilization of a linearly decreasing inertia weight, and PSO may fail to find the required optima when the problem to be solved is too complicated and complex. K-means is the most widely used and studied clustering algorithm. Given a set of n data points in real d-dimensional space, Rd , and an integer k, the clustering problem is to determine a set of k points in Rd , the set of k points is called cluster centres, and the set of n data points is divided into k groups according to the distance between it and cluster centres. K-means algorithm is simple and flexible, but it has some shortcomings, the cluster result is sensitive to the selection of the initial cluster centroids and may converge to the local optima. However, the same initial cluster centre in a data space can always generate the same cluster results, if a good cluster centre can always be obtained, the K-means will work well. K-mean clusteringThe term " k-means" was first used by James MacQueen in 1967. The standard algorithm was first proposed by Stuart Lloyd in 1957 as a technique for pulse-code modulation, though it wasn't published until 1982.

The K-Means clustering algorithm is a partition-based cluster analysis

method. According to the algorithm we firstly select k objects as initial cluster centers, then calculate the distance between each object and each cluster center and assign it to the nearest cluster, update the averages of all clusters, repeat this process until the criterion function converged. Square error criterion for clustering, Xij is the sample j of i-class, mi is the center of i-class, ni is the number of samples i-class, Algorithm step are shown in the fig(1). K-means clustering algorithm is simply described as follows: Input: N objects to be cluster {x1, x2..... xn}, the number of clusters k; Output: k clusters and the sum of dissimilarity betweeneach object and its nearest cluster center is the smallest; Arbitrarily select k objects as initial cluster centers(m1, m2... mk); Calculate the distance between each object xi andeach cluster center, then assign each object to the nearestcluster, formula for calculating distance as: 2i = 1, 2..... Nj = 1, 2..... kd (xi, mj) is the distance between data i and cluster j; Calculate the mean of objects in each cluster as thenew cluster centers, i= l, 2 . . . k; Ni is the number of samples of current cluster i; No. of Cluster KCentroidCalculate distance betweenObject and CentroidMake group base onminimum distanceObjectmove togroup? OutputFig(1). Flowchart of K-meanParticle swarm optimizationPSO was introduced by Kennedy and Eberhart, it was inspired by the swarming behavior of animals and humansocial behavior. A particle swarm is a population of particles, in which each particle is a moving object which can move through the search space and can be attracted to the better positions. PSO must have a fitness evaluation function to decide the better and best positions, the function can take the particle’s position and assigns it a fitness value. Then the objective is to optimize the fitness function. In general, the fitness function is pre-defined and is depend on the

problem. Each particle has own coordinate and velocity to change the flying direction in the search space. And all particles move through the search space by following the current optimum particles. Each particle consists of a position vector z, which can represent the candidate solution to the problem, a velocity vector v, and a memory vector pid, which is the better candidate solution encountered by a particle. Suppose the search space is n-dimensional, then the ith individual can be represented as: Zi = {Zi1, Zi2, ……. Zin}Vi = {Vi1, Vi2,……. Vin}i= 1, 2, 3…….. n. Where n is the size of swarm. The best previous experience of the ith particle is represented as: pid_i = {pid_i1, pid_i2……. pid_in}Another memory vector pgd is used, which is the best candidate solutions encountered by all particles. The particles are then manipulated according to the following equations: V_id (t+1)= wv_id (t)+η1rand(pid_i-Z_id (t))+η2 rand(pgd-Z_id (t)), Z_id (t+1)= Z_id (t)+V_id (t+1), d= 1, 2…. nWhere w is an inertia weight, which used to control the impact of the previous history of velocities on the current velocity, and regulate the trade-off between the global and local exploration abilities of the swarm. A big inertia weight facilitates global exploration, while a small one tends to facilitate local exploration. In order to get a better global exploration, w can be gradually decreased to get a better solution. η1 and η2 are two positive constants, rand is a uniformly generated random number. The equation shows that in calculating the next velocity for a particle, the previous velocity of the particle, the best location in the neighborhood about the particle, the global best location all contribute some influence to the next velocity. Particle’s velocities in each dimension can arrive to a maximum velocity vmax, which is defined to the range of the search space in each dimension. The process of the PSO can be described as

follows: Firstly, initialize a population of particles with random positions and velocities in the search space. Secondly, for each particle i, update the position and velocity according to , compute the fitness value according to the fitness function, update pid_i and pgd if necessary, repeat this process until termination conditions are met. NoYesTrainee DataInitialize PopulationUpdate pid_iUpdate pgdUpdate velocityUpdate PositionConditionMet? outputAdvantage & disadvantage of K-mean and psoAdvantages of K-mean clusteringK-mean clustering is simple and flexible. K-mean clustering algorithm is easy to understandand implements. Disadvantages of K-mean clusteringIn K-mean clustering user need to specify the numberof cluster in advanced. K-mean clustering algorithm performance dependson a initial centroids so the algorithm gives no guarantee for an optimal solution. Advantages of PSOPSO based on the intelligence and it is applied onboth scientific research and engineering. PSO have no overlapping and mutation calculation. The search can be carried out by the speed of the particle. During the development of several generations, only the most optimist particle can transmit information onto the other particles, and the speed of the researching is very fast. PSO adopts the real number code, and it is decideddirectly by the solution. calculation in PSO very simple and it is efficient in global search. Disadvantages of PSOIt is slow convergence in refined search stage andweak local search ability. The method cannot work out the problems of non-coordinate system, such as the solution to the energy field and the moving rules of the particles in the energy field. ConclusionsStudy of the k-mean clustering and Particle swam optimization we say that the k-mean which is depend on initial condition, which may cause the algorithm converge to suboptimal solution. On the

other side Particle swarm optimization is less sensitive for initial condition due to its population based nature. So Particle swarm optimization is more likely to find near optimal solution. AcknowledgmentPritesh vora wish to acknowledge and thanks Asst. Prof. Bhavesh Oza, for him guidance and help for doing this work. He also acknowledges Prof. D. A. Parikh, Head of computer department, and to all staff of computer department for completion of this work.