© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 08 Issue: 01 | January - March 2017
Paper is available at www.jrps.in | Email : info@jrps.in

A REVIEW-DENSITY BASED CLUSTERING ANALYSIS USING NEURAL NETWORK

Poonam Malik[1], Kirti Gautam[2]
[1]Research Scholar, JIET, Jind , [2]Assistant Professor, JIET, JIND,

ABSTRACT: Analysis of data sets can find new correlations to spot business trends, prevent diseases, combat crime and so on. Scientists, business executives, practitioners of medicine, advertising and governments alike regularly meet difficulties with large data sets in areas including Internet search, finance and business informatics. Scientists encounter limitations in e-Science work, including meteorology, genomics, connectomics, complex physics simulations, biology and environmental research. Data sets are growing rapidly in part because they are increasingly gathered by cheap and numerous information-sensing mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks.

[1] INTRODUCTION

Density Based clustering

Density-based spatial clustering of applications with noise (DBSCAN) is a data clustering algorithm proposed by Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu in 1996.[1] It is a density-based clustering algorithm: given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors), marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away). DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature Clustering is the process of making a group of abstract objects into classes of similar objects. A cluster of data objects can be treated as one group. While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups. The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

Neural Network

Artificial neural networks are typically specified using three things Architecture specifies what variables are involved in the network and their topological relationships—for example the variables involved in a neural network might be the weights of the connections between the neurons, along with activities of the neurons Activity Rule Most neural network models have short time-scale dynamics: local rules define how the activities of the neurons change in response to each other. Typically the activity rule depends on the weights (the parameters) in the network.

Learning Rule The learning rule specifies the way in which the neural network's weights change with time. This learning is usually viewed as taking place on a longer time scale than the time scale of the dynamics under the activity rule. In machine learning and cognitive science, an artificial neural network (ANN) is a network inspired by biological neural networks (the central nervous systems of animals, in particular the brain) which are used to estimate or approximate functions that can depend on a large number of inputs that are generally unknown.

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 08 Issue: 01 | January - March 2017
Paper is available at www.jrps.in | Email : info@jrps.in

Employing artificial neural networks

Perhaps greatest advantage of ANNs is their ability to be used as an arbitrary function approximation mechanism that 'learns' from observed data.[42] However, using them is not so straightforward, & a relatively good understanding of underlying theory is essential.

- Choice of model: This would depend on data representation & application. Overly complex models tend to lead to challenges in learning.
- Learning algorithm: There is numerous trades-offs between learning algorithms. Almost any algorithm would work well within *correct hyper parameters* for training on a particular fixed data set. However, selecting & tuning an algorithm for training on unseen data require a significant amount of experimentation.
- Robustness: If model, cost function & learning algorithm are selected appropriately, resulting ANN could be extremely robust.

With correct implementation, ANNs could be used naturally in online learning & large data set applications. Their simple implementation & existence of mostly local dependencies exhibited in structure allows for fast, parallel implementations in hardware.

[2] CLUSTERING METHODS

Clustering methods can be classified into the following categories, Partitioning Method, Hierarchical Method, Density-based Method, Grid-Based Method, Model-Based Method, and Constraint-based Method. Partitioning Method Suppose we are given a database of 'n' objects and the partitioning method constructs 'k' partition of data. Each partition will represent a cluster and $k \leq n$. It means that it will classify the data into k groups, which satisfy the following requirements − Each group contains at least one object. Each object must belong to exactly one group. Hierarchical Methods, this method creates a hierarchical decomposition of the given set of data objects. We can classify hierarchical methods on the basis of how the hierarchical decomposition is formed. There are two approaches here, Agglomerative Approach and Divisive Approach

[3] DESIGM METHODOLOGY

Design Methodology Data mining is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful & understandable patterns in large databases. patterns must be actionable so that they may be used in an enterprise's decision making process. It is usually used by business intelligence organizations, & financial analysts, but this is increasingly used in sciences to extract information from enormous data sets generated by modern experimental & observational methods. A typical example for a data mining scenario may be "In context of a super market, if a mining analysis observes that people who buy pen tend to buy pencil too, then for better business results seller could place pens & pencils together." Data mining strategies could be grouped as follows: • **Classification**- Here given data instance has to be classified into one of target classes which are already known or defined [19, 20]. One of examples could be whether a customer has to be classified as a trustworthy customer or a defaulter with in a credit card transaction data base, given his various demographic & previous purchase Characteristics.

• **Estimation**- Like classification, purpose of an estimation model is to determine a value for an unknown output attribute. However, unlike classification, output attribute for an estimation problem are numeric rather than categorical.

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 08 Issue: 01 | January - March 2017
Paper is available at www.jrps.in | Email : info@jrps.in

• **Prediction**- It is not easy to differentiate prediction from classification or estimation. only difference is that rather than determining current behaviour, predictive model predicts a future outcome. output attribute could be categorical or numeric. An example could be "Predict next week's closing price for Dow Jones Industrial Average". [53, 65] explains construction of a decision tree & its predictive applications.

• **Association rule mining** -Here interesting hidden rules called association rules with in a large transactional data base is mined out. For e.g. rule {milk, butter->biscuit} provides information that whenever milk & butter are purchased together biscuit is also purchased, such that these items could be placed together for sales to increase overall sales of each of items.

[4] SOFT COMPUTING WITH CLUSTERING

Soft computing is an emerging approach to computing which parallels remarkable potential of human mind to reason & learn within an environment of uncertainty & imprecision[12].Soft Computing consists of several computing paradigms like Neural Networks, Fuzzy Logic, & Genetic algorithms. Soft Computing uses hybridization of these techniques. A hybrid technique would inherit all benefits of constituent techniques. Thus elements of Soft Computing are complementary, not competitive, offering their own advantages & techniques to partnerships to allow solutions to otherwise unsolvable problems. Recently various soft computing methodologies have been applied to handle different challenges posed by data mining. Soft computing methodologies (involving fuzzy sets, neural networks, genetic algorithms, & rough sets) are most widely used within data mining step of overall KDD process. Fuzzy sets provide a natural framework for process within dealing with uncertainty. Neural networks & rough sets are widely used for classification & rule generation. Genetic algorithms (GAs) are involved within various optimization & search processes, like query optimization & template selection. Other approaches like case based reasoning [5] & decision trees [12], [13] are also widely used to solve data mining problems. Each of them contributes a distinct methodology for addressing problems within its domain. This is done within a cooperative, rather than a competitive, manner. The result is a more intelligent & robust system providing a human-interpretable, low cost, approximate solution, as compared to traditional techniques.

[5] PROPOSED WORK

Density Based clustering

The most popular density based clustering method is DBSCAN. In contrast to many newer methods, it features a well-defined cluster model called "density-reachability". Similar to linkage based clustering, it is based on connecting points within certain distance thresholds. However, it only connects points that satisfy a density criterion, in original variant defined as a minimum number of other objects within this radius. A cluster consists of all density-connected objects (which could form a cluster of an arbitrary shape, in contrast to many other methods) plus all objects that are within these objects' range. Another interesting property of DBSCAN is that its complexity is fairly low - it requires a linear number of range queries on database - & that it will discover essentially same results (it is deterministic for core & noise points, but not for border points) in each run, therefore there is no need to run it multiple times. OPTICS is a generalization of DBSCAN that removes

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 08 Issue: 01 | January - March 2017
Paper is available at www.jrps.in | Email : info@jrps.in

need to choose an appropriate value for range & produces a hierarchical result related to that of linkage clustering. DeLi-Clu, Density-Link-Clustering combines ideas from single-linkage clustering & OPTICS, eliminating parameter entirely & offering performance improvements over OPTICS by using an R-tree index. The key drawback of DBSCAN & OPTICS is that they expect some kind of density drop to detect cluster borders. Moreover, they cannot detect intrinsic cluster structures which are prevalent in majority of real life data. A variation of DBSCAN, EnDBSCAN,[14] efficiently detects such kinds of structures. On data sets with, for example, overlapping Gaussian distributions - a common use case in artificial data - cluster borders produced by these algorithms will often look arbitrary, because cluster density decreases continuously. On a data set consisting of mixtures of Gaussians, these algorithms are nearly always outperformed by methods such as EM clustering that are able to precisely model this kind of data. Mean-shift is a clustering approach where each object is moved to densest area in its vicinity, based on kernel density estimation. Eventually, objects converge to local maxima of density. Similar to k-means clustering, these "density attractors" could serve as representatives for data set, but mean-shift could detect arbitrary-shaped clusters similar to DBSCAN. Due to expensive iterative procedure & density estimation, mean-shift is usually slower than DBSCAN or k-Means.

Performance Improvement of Web Usage Mining By Using Learning Based Density Based Clustering

Due to increasing amount of data available online, World Wide Web has becoming one of most valuable resources for information retrievals & knowledge discoveries. Web mining technologies are right solutions for knowledge discovery on Web. Knowledge extracted from Web could be used to raise performances for Web information retrievals, question answering, & Web based data warehousing. In present work, we propose a new technique to enhance learning capabilities & reduce computation intensity of a competitive learning multi-layered neural network using Density Based clustering algorithm.

[6] CONCLUSION

The advent of laptops, palmtops, cell phones, & wearable computers is making ubiquitous access to huge quantity of information possible. Advanced analysis of data for extracting useful knowledge is next natural step within world of ubiquitous computing. Accessing & analyzing data from a ubiquitous computing device offer many challenges.

Before data mining algorithms could be used, a target data set must be assembled. As data mining could only uncover patterns actually present within data, target data set must be large sufficient to contain these patterns while remaining concise enough to be mined within an acceptable time limit. A common source of data is considered data mart or data warehouse. Pre-processing is essential to analyze multivariate data sets before data mining. The target set is then cleaned. We have described the substantial technical challenges in developing and deploying decision support systems. While many commercial products and services exist, there are still several interesting avenues for research. We will only touch on a few of these here.

REFERENCE

1. Hong Liu 1, and Xiaohong Yu(2009) wrote on Application Research of k-means Clustering Algorithm in Image Retrieval System

2. Dr. Yashpal singh, 2alok singh chauhan in 2009 neural networks in data mining
3. In 2011 Jiawei Han & Jing Gao University of Illinois at Urbana-Champaign wrote paper on "Research Challenges for Data Mining in Science & Engineering"
4. Navjot Kaur, Jaspreet Kaur Sahiwal, Navneet Kaur (2012) wrote on Efficient K-Means Clustering Algorithm Using Ranking Method In Data Mining
5. Manjot Kaur * Navjot Kaur 2012 Web Document Clustering Approaches Using K-Means Algorithm
6. Prabin Lama 2013 Clustering system based on text mining using k-means algorithm
7. Farhat Roohi in 2013 Artificial Neural Network Approach to Clustering
8. Waldemar Wójcik & Konrad Gromaszek in 2014 (Lublin University of Technology, Poland) introduced "Data Mining Industrial Applications".
9. Piatetsky-Shapiro, Gregory (1991), Discovery, analysis, and presentation of strong rules, in Piatetsky-Shapiro, Gregory; and Frawley, William J.; eds., Knowledge Discovery in Databases, AAAI/MIT Press, Cambridge, MA.
10. Agrawal, R.; Imieliński, T.; Swami, A. (1993). "Mining association rules between sets of items in large databases". Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93. p. 207. doi:10.1145/170035.170072. ISBN 0897915925.
11. Hahsler, Michael (2005). "Introduction to arules – A computational environment for mining association rules and frequent item sets" (PDF). Journal of Statistical Software.
12. Michael Hahsler (2015). A Probabilistic Comparison of Commonly Used Interest Measures for Association Rules. http://michael.hahsler.net/research/association_rules/measures.html
13. Hipp, J.; Güntzer, U.; Nakhaeizadeh, G. (2000). "Algorithms for association rule mining --- a general survey and comparison". ACM SIGKDD Explorations Newsletter **2**: 58. doi:10.1145/360402.360421.
14. Tan, Pang-Ning; Michael, Steinbach; Kumar, Vipin (2005). "Chapter 6. Association Analysis: Basic Concepts and Algorithms" (PDF). Introduction to Data Mining. Addison-Wesley. ISBN 0-321-32136-7.
15. Pei, Jian; Han, Jiawei; and Lakshmanan, Laks V. S.; Mining frequent itemsets with convertible constraints, in Proceedings of the 17th International Conference on Data Engineering, April 2–6, 2001, Heidelberg, Germany, 2001, pages 433-442
16. Agrawal, Rakesh; and Srikant, Ramakrishnan; Fast algorithms for mining association rules in large databases, in Bocca, Jorge B.; Jarke, Matthias; and Zaniolo, Carlo; editors, Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), Santiago, Chile, September 1994, pages 487-499

17. Zaki, M. J. (2000). "Scalable algorithms for association mining". IEEE Transactions on Knowledge and Data Engineering **12** (3): 372–390. doi:10.1109/69.846291.

18. Hájek, Petr; Havel, Ivan; Chytil, Metoděj; The GUHA method of automatic hypotheses determination, Computing 1 (1966) 293-308

19. Hájek, Petr; Feglar, Tomas; Rauch, Jan; and Coufal, David; The GUHA method, data preprocessing and mining, Database Support for Data Mining Applications, Springer, 2004, ISBN 978-3-540-22479-2