International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 9, Issue 5 (December 2013), PP. 23-26

A Review on Clustering Based Methods and Usage for Pattern Recognition

Deepti Mishra
Department of computer science and engineering
I.T.S. Engineering College, Greater Noida

Abstract:- Pattern recognition is an important field of computer science concerned with recognizing patterns, particularly visual and sound patterns. It is an area of work that has applications in many disciplines. Pattern recognition is the science of making inferences based on data. It is the study of how machines can observe the environment, learn to distinguish patterns of interest, make sound and reasonable decisions about the categories of the patterns. The pattern recognition is based on the concept of supervised learning and unsupervised learning. The approach based on supervised learning is called classification and the approach based on unsupervised learning is called clustering. Both the approaches are the techniques of data mining. Data mining is the approach which searches for new, valuable and nontrivial information from large volumes of data. The study includes the techniques that are based on clustering and are useful for pattern recognition. The study aims at providing the review of clustering techniques and their applications in pattern recognition. The discussion on the study will guide the researchers for improving their research direction.

Keywords:- Pattern Recognition, Data Mining, Clustering, unsupervised learning, supervised learning.

I. INTRODUCTION

In day to day life a person goes through an endless number of pattern recognition problems such as smells, images, voices, faces, situations, and so on. Most of these problems we solve at a sensory level or intuitively, without an explicit method or algorithm. As soon as we are able to provide an algorithm the problem becomes trivial and we happily delegate it to the computer. Indeed, machines have confidently replaced humans in many formerly difficult or impossible way, now just tedious pattern recognition tasks such as mail sorting, medical test reading, military target recognition, signature verification, meteorological forecasting, DNA matching, fingerprint recognition, and so on. Pattern Recognition (PR) is the scientific discipline whose goal is the classification of data, objects or, in general, patterns into categories or classes. This involves capabilities of – unsupervised learning – supervised learning [13].

Classification and clustering both are the important approaches of pattern recognition [5]. Classification is based on supervised learning. In supervised learning the class label of each training sample is provided. Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be “the process of organizing objects into groups whose members are similar in some way”. A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. Clustering techniques identify the clusters into which the data can be divided; the similarity criterion is distance: two or more objects belong to the same cluster if they are “close” according to a given distance (in this case geometrical distance). This is called distance-based clustering [1]. Another kind of clustering is conceptual clustering: two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

Clustering is applicable in many areas including pattern detection, Data mining, Information retrieval, text mining, Web analysis, marketing, medical diagnostic and many more [6]. Here the study includes pattern recognition and its techniques those are based on clustering. Clustering is a statistical tool**,** which helps discovering the relative importance of various dimensions in defining the belongings of particular item to a particular category.

II. RELATED WORK

Pattern recognition has become a very important field over the last decade since automation and computerization in many systems has led to large amount of data being stored in the databases. The primary intention of pattern recognition is to automatically assist humans in analyzing the vast amount of available data and extracting useful knowledge from it. Many algorithms have been developed for many applications, especially for static pattern recognition [10].

Clustering has its roots in many areas including data mining, statistics and pattern recognition [6]. By clustering, one can identify dense and sparse regions and, therefore, discover overall distribution patterns and interesting correlation among data attributes [6].

A new parametric approach has been proposed, which starts with an estimate of the local distribution and efficiently avoids pre-assuming the cluster number. This clustering program is applied to both artificial and benchmark data classification and its performance is proven better than the well-known k-means algorithm [9]. Research work has presented the Clustered Clause structure, which uses information-based clustering and dependencies between sentence components to provide a simplified and generalized model of a grammatical clause. Research work has done, which is based on dependencies within the sentence, enables us to detect complex textual relations at a higher level of context. The relations we detect are of interest in themselves, as linguistic phenomena, and are also highly suited for use in certain linguistic and cognitive tasks [7].

Clustering is the process of grouping of data, where the grouping is established by finding similarities between data based on their characteristics [12]. Such groups are termed as Clusters. A comparative study of clustering algorithms across two different data items has been done. The performance of the various clustering algorithms is compared based on the time taken to form the estimated clusters. The experimental results of various clustering algorithms to form clusters can be depicted as a graph. Thus it can be concluded as the time taken to form the clusters increases such as the number of clusters increases. The farthest first clustering algorithm takes very few seconds to cluster the data items whereas the simple K-Means takes the longest time to perform clustering [3].

A simple and efficient implementation of Lloyd's k-means clustering algorithm has been presented in a reference, which we call the filtering algorithm. This algorithm is easy to implement, requiring a kd-tree as the only major data structure. It has established the practical efficiency of the filtering algorithm in two ways. First, it presents a data-sensitive analysis of the algorithm's running time, which shows that the algorithm runs faster as the separation between clusters increases. Second, it presents a number of empirical studies both on synthetically generated data and on real data sets from applications in color quantization, data compression, and image segmentation.

Research has been done with the artificial intelligence and pattern recognition i.e. how to teach a computer to distinguish relevant signals from noise and how to use this information to make decisions [2].

III. METHODS and TECHNIQUES

Pattern recognition is about assigning labels to objects. Objects are described by a set of measurements called also attributes or features. If the data set is not given, an experiment is planned and a data set is collected. The relevant features have to be nominated and measured. Pattern recognition is generally categorized according to the type of learning procedure used to generate the output value. Supervised learning assumes that a set of training data (the training set) has been provided, consisting of a set of instances that have been properly labeled by hand with the correct output. A learning procedure then generates a model that attempts to meet two, sometimes conflicting, objectives: Perform as well as possible on the training data, and generalize as well as possible to new data. Unsupervised learning, on the other hand, assumes training data that has not been hand-labeled, and attempts to find inherent patterns in the data that can then be used to determine the correct output value for new data instances. Pattern recognition is the important application of clustering. The clustering is a set of techniques for classification of samples into a number of groups. Therefore, the samples in one group are grouped and samples belonging to different groups are grouped as another group. The input of clustering is a set of samples and the process of clustering is to measure the similarity and or dissimilarity between given samples. The output of the clustering is a number of groups or clusters in the form of graphs, histograms and normal computer results showing group number.

Clustering algorithms may be classified as Partitioning Clustering, Density Based Clustering, Hierarchical clustering [4][12].

Partitioning Clustering: The basic concept in partitioning clustering is to divide the data into proper subset in recursive manner, and go through each subset and relocate points between clusters. Given a database of n objects, a partitioning method constructs k(n) partitions of the data, where each partition represents a cluster. That is, it classifies the data into k groups, which together satisfy the following requirements:

Each group must contain at least one object and each object must belong to exactly one group. K-means is the basic technique of partitioning clustering which is based on calculating the Euclidean distance.

Hierarchical clustering: Use distance matrix as clustering criteria. This method does not require the number of clusters ***k*** as an input, but needs a termination condition. It works by grouping data objects into tree of clusters. They are either agglomerative (bottom-up) or divisive (top-down):

(a) Agglomerative algorithms start with each object being a separate cluster itself, and successively merge groups according to a distance measure. The clustering may stop when all objects are in a single group or at any other point the user wants.

These methods generally follow a greedy-like bottom-up merging. Chameleon is the basic algorithm for agglomerative clustering which merges all the set of points into a sparse graph and then partition them into clusters. CURE is also the important approach for it.

(b) Divisive algorithms follow the opposite strategy. They start with one group of all objects and successively split groups into smaller ones, until each object falls in one cluster, or as desired.

Divisive approaches divide the data objects in disjoint groups at every step, and follow the same pattern until all objects fall into a separate cluster. This is similar to the approach followed by divide-and-conquer algorithms.

Density Based clustering: The general idea is to continue growing the given cluster as long as the density (number of objects or data points) in the "neighborhood" exceeds some threshold i.e. for each data point within a cluster, the neighborhood of a given radius has to contain at least a minimum number of points. Such a method is used to filter out noise i.e. outliers; and discover clusters of arbitrary shapes. The basic approach used is DBSCAN and OPTICS algorithms. DBSCAN algorithm grows regions with sufficiently high density into clusters.

There are two major approaches to pattern recognition, which are: numerical and syntactic Approach [11]. The former one can be further divided into different fields as shown in Figure 1.

Stages in clustering for pattern recognition

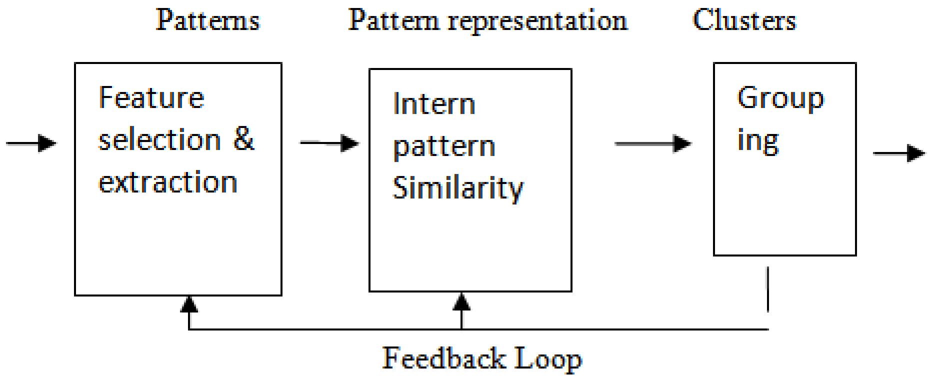

Figure 1

Data is collected both to train and to test the system. The selection of the distinguished features is a typical design step. Prior knowledge plays a critical role. The process of using data to determine the classifier is referred to as training the classifier. That trained classifier is used to define the clusters. Those patterns which are similar to one another are kept in same group but those are different from patterns kept in other groups. The loop iterates itself until the resulted pattern is calculated.

IV. CONCLUSION

Tremendous volumes of data are filled in the computers and in the internet. The Government agencies, scientific institutions and businesses have all dedicated enormous resources to collecting and storing data. In the real world, only a small amount of these data is used due to the fact that volumes are simply too large to manage, the data structures themselves are too complicated to be analyzed effectively. Pattern recognition techniques are concerned with the theory and algorithms of putting abstract objects, e.g., measurements made on physical objects, into categories. Typically the categories are assumed to be known in advance, although there are techniques to learn the categories (clustering). In clustering there is no explicit teacher, and the system forms clusters of the input patterns. Input patterns are not labeled.

REFERENCES

[1]. Pratima D. and Nimmakanti N., " Pattern Recognition algorithms for cluster identification problem", Special Issue of International Journal of Computer Science & Informatics (IJCSI),vol.- II, Issue-1, 2.

[2]. Richard j. Morris, "Statistical pattern recognition for macromolecular crystallographers", Acta crystallographica, Biological crystallography, ISSN-0907-4449, 2004, D60, 2133-2143.

[3]. S.Revathi and Dr. T. Nalini, "Performance comparison of various clustering algorithm", ijarcsse, vol 3, issue 2, Feb 2013, pg 67-72.

[4]. Koteeswaran S., P.Visu and J.Janet, "A review on clustering and outlier analysis techniques in data mining", American journal of applied sciences 9(2), ISSN 1546-9239, 2012, pg 254-258.

[5]. Richard o. Duda, Peter E. Hart, David G. Stork, “ Pattern Classification”,

[6]. Jiawei Han, M. Kamber, “ Data Mining: Concepts and techniques”.

[7]. Brody S., “Cluster based Pattern recognition in natural language text”, aug 2005.

[8]. Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and

[9]. Angela Y. Wu, “An efficient K-means clustering algorithm: analysis and implementation”, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 24, NO. 7, JULY 2002.

[10]. Zeng Y. and Starzyk J., “ Statistical approach to clustering in pattern recognition”,IEEE 2001.

[11]. Sadina Gagula-Palalic, “Fuzzy clustering models and algorithms for pattern recognition”.

[12]. Jain A.K., Murty M.N. and Flynn P.J., “Data Clustering : A review”, ACM Computing Surveys, Vol. 31, No. 3, Sep 1999.

[13]. Gupta A., Gupta A. and Mishra A., “Research paper on clustering techniques of data variations”, IJATER, vol 1, 2011.

[14]. Pankaj Sharma, “Pattern Recognition”.

[15]. Jain A.K. and Maheshwari S., “ Survey of clustering techniques in data mining”, IJCSMR, Aug 2012.