www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

"Clustering Based Data Ensemble Using Incremental Semi Supervised Clustering Method"

Mrs.G.KALAIMAGAL

PG-Scholar

Department of Computer Science and Engineering

P.S.V College of Engineering and Technology

Karhsnigiri-635108, Tamilnadu, India

Prof.B.SAKTHIVEL, M.E,

Professor & Head

Department of Computer Science and Engineering

P.S.V College of Engineering and Technology

Karhsnigiri-635108, Tamilnadu, India

Mail id : sakthi_mnnit@yahoo.co.in

Mobile no : 9787447671

ABSTRACT

Traditional cluster ensemble approaches have three limitations: (1) They do not make use of prior knowledge of the datasets given by experts. (2) Most of the conventional cluster ensemble methods cannot obtain satisfactory results when handling high dimensional data. (3) All the ensemble members are considered, even the ones without positive contributions.

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

In order to address the limitations of conventional cluster ensemble approaches, we first propose an incremental semi-supervised clustering ensemble framework (ISSCE) which makes use of the advantage of the random subspace technique, the constraint propagation the proposed incremental ensemble member selection process, and the normalized cut algorithm to perform high dimensional data clustering. The incremental ensemble member selection process is newly designed to judiciously remove redundant ensemble members based on a newly proposed local cost function and a global cost function, and the normalized cut algorithm is adopted to serve as the consensus function for providing more stable, robust and accurate results. Then, a measure is proposed to quantify the similarity between two sets of attributes, and is used for computing the local cost function in ISSCE. Next, we analyze the time complexity of ISSCE theoretically. Finally, a set of nonparametric tests are adopted to compare multiple semi-supervised clustering ensemble approaches over different datasets.

Key words: cluster, satisfactory, semi-supervised, proposed, nonparametric

CHAPTER 1

1.1. INTRODUCTION

Recently, cluster ensemble approaches are gaining more and more attention [1]-[4], due to its useful applications in the areas of pattern recognition [2]-[5], data mining [6][7], bioinformatics [8]-[10], and so on. When compared with traditional single clustering algorithms, cluster ensemble approaches are able to integrate multiple clustering solutions obtained from different data sources into a unified solution, and provide a more robust, stable and accurate final result. However, conventional cluster ensemble approaches have several limitations: (1) They do not consider how to make use of prior knowledge given by

Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

experts, which is represented by pairwise constraints. Pairwise constraints are often defined as the must-link constraints and the cannot-link constraints. The must-link constraint means that two feature vectors should be assigned to the same cluster, while the cannot-link constraints means that two feature vectors cannot be assigned to the same cluster. (2) Most of the cluster ensemble methods cannot achieve satisfactory results on high dimensional datasets.

Not all the ensemble members contribute to the final result. In order to address the first and second limitations, we first propose the random subspace based semi-supervised clustering ensemble framework (RSSCE), which integrates the random subspace technique [11], the constraint propagation approach [12], and the normalized cut algorithm [13] into the cluster ensemble framework to perform high dimensional data clustering. Then, the incremental semi-supervised clustering ensemble framework (ISSCE) is designed to remove the redundant ensemble members. When compared with traditional semi-supervised clustering algorithm, ISSCE is characterized by the incremental ensemble member selection process based on a newly proposed global objective function and a local objective function, which selects ensemble members progressively.

The local objective function is calculated based on a newly designed similarity function which determines how similar two sets of attributes are in the subspaces. Next, the computational cost and the space consumption of ISSCE are analyzed theoretically. Finally, we adopt a number of nonparametric tests to compare multiple semi-supervised clustering ensemble approaches over different datasets. The experiment results show the improvement of ISSCE over traditional semi supervised clustering ensemble approaches or conventional cluster ensemble methods on 6 real-world datasets from UCI machine learning repository

Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

[14] and 12 real-world datasets of cancer gene expression profiles. The contributions of the paper is fourfold. First, we propose an incremental ensemble framework for semi-supervised clustering in high dimensional feature spaces. Second, a local cost function and a global cost function are proposed to incrementally select the ensemble members.

Third, the newly designed similarity function is adopted to measure the extent to which two sets of attributes are similar in the subspaces. Fourth, we use non-parametric tests to compare multiple semi-supervised clustering ensemble approaches over different datasets. The remainder of the paper is organized as follows. Section II describes previous work related to semi-supervised clustering and cluster ensemble. Section III presents the incremental semi-supervised clustering ensemble framework. Section IV analyzes the proposed algorithm theoretically. Section V experimentally evaluates the performance of our proposed approach. Section VI describes our conclusion and future work.

Cluster ensemble, also referred to as consensus clustering, is one of the important research directions in the area of ensemble learning, which can be divided into two stages: the first stage aims at generating a set of diverse ensemble members, while the objective of the second stage is to select a suitable consensus function to summarize the ensemble members and search for an optimal unified clustering solution. To attain these objectives, Strehl et al. [1] first proposed a knowledge reuse framework which integrates multiple clustering solutions into a unified one. After that, a number of researchers followed up Strehl's work, and proposed different kinds of cluster ensemble approaches [15]-[21]. While there are different kinds of cluster ensemble techniques, few of them consider how to handle high dimensional data clustering, and how to make use of prior knowledge. High dimensional datasets have too many attributes relative to the number of samples, which will

lead to the overfitting problem. Most of the conventional cluster ensemble methods do not take into account how to handle the overfitting problem, and cannot obtain satisfactory results when handling high dimensional data. Our method adopts the random subspace technique to generate the new datasets in a low dimensional space, which will alleviate this problem. There are also other research works which study the properties of the cluster ensemble theoretically, such as the stability of k-means based cluster ensemble [2], the efficiency of the cluster ensemble [22], the convergence property of consensus clustering [23], the scalability property of the cluster ensemble [24], the effectiveness of cluster ensemble methods [25], and so on.

Cluster ensemble approaches have been applied to different areas, such as bioinformatics [26][27], image segmentation [28], language processing [29], Internet security [30], and so on. Recently, some researchers realized that not all the ensemble members contribute to the final result, and investigate how to select a suitable subset of members to obtain better results [31]-[35]. For example, Yu et al. [33]-[34] treated the ensemble members as features, and explored how to use suitable feature selection techniques to choose the ensemble members. In summary, most of the cluster ensemble approaches only consider using a similarity score or feature selection technique to remove the redundant ensemble members, and few ofthem study how to apply an optimization method to search for a suitable subset of ensemble members.

In the current work, the proposed ISSCE framework uses a newly designed incremental ensemble member selection process to generate an optimal set of members. In addition, conventional cluster ensemble methods do not take into account how to make use of prior knowledge, which is usually represented in the form of pairwise constraints or a

very small set of labeled data. Single semi-supervised clustering algorithms have the ability to handle prior knowledge, and use them to guide the search in the process of clustering. A number of semi-supervised clustering algorithms have been proposed [36]-[45], such as semi-supervised maximum margin clustering [36], semi-supervised kernel mean shift clustering [37], semi-supervised linear discriminant clustering [38], semisupervised hierarchical clustering [39], active learning based semi-supervised clustering [40], semi-supervised affinity propagation [41], semi-supervised nonnegative matrix factorization [42], and so on. It is natural to adopt a suitable single semisupervised clustering method as the basic clustering algorithm in the cluster ensemble. In this paper, we consider the constraint propagation approach (E2CP) proposed in [12], which propagate constraints in a more exhaustive and efficient way, as the basic clustering algorithm in ISSCE.

This approach has two advantages: (1) The time complexity of E2CP is proportional to the total number of all possible pairwise constraints, which is O(Kn2) (where K is the number of neighbors in the K-NN graph, and n is the number of feature vectors in the dataset). It is much smaller than that of conventional constraint clustering approaches, which is O(n 4). (2) E 2CP achieves good results on different real-world datasets, such as image datasets, UCI datasets, cross-modal multimedia retrieval, and so on. Greene and Cunningham [55] studied constraint selection by identifying the constraints which are useful for improving the accuracy of the clustering solution. When compared with their work, our proposed incremental semi-supervised clustering ensemble framework adopts the more effective constraint propagation approach to convey supervised information from the labeled data samples to the unlabeled samples, and solve the label propagation problem in parallel.

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

1.2 EXISTING SYSTEM

Semi-supervised clustering is an attractive alternative for traditional (unsupervised) clustering in targeted applications. By using the information of a small annotated dataset, semi-supervised clustering can produce clusters that are customized to the application domain. In this paper, we present a semi-supervised clustering technique based on a multi-objective evolutionary algorithm (NSGA-II-clus). We apply this technique to the task of clustering medical publications for Evi-dence Based Medicine (EBM) and observe an improvement of the results against unsupervised and other semi-supervised clustering techniques.

1.3 PROPOSED SYSTEM

In our present work we propose to develop a semi-supervised clustering technique and apply that for EBM. The proposed approach uses only 10% labeled information which is easy to obtain. The proposed technique is novel in a way that it uses the labeled information during the internal steps of the proposed clustering process. More specifically we can say that the internal steps of NSGA-II based clustering are modified to take care of this labeled information. The labeled information was used by Ekbal et al. (2013) to select a single solution from the final Pareto optimal front after the execution of AMOSA based clustering technique. Thus, the use of NSGA-II as the underlying optimization technique makes the system less complex and time consuming. In this paper, we propose the use of NSGA-II (Deb et al., 2013) for semi-supervised clustering of documents. We propose two different versions of the NSGA-II based semi-supervised clustering technique. In the first approach the available supervised information in the form of must-link and cannot-link

constraints can be used during the selection phase of clustering. These constraints are taken into account while calculating crowding distance which is further used to assign ranks to different solutions of the combined population.Thus, the available supervised information is used in each generation of the proposed technique. In the second approach, we use a semi-supervised approach to select a single solution from the set of final solutions produced by the MOO-based approach. In this case, supervised information is used only at the final stage rather than during the clustering phase.

CHAPTER 2

WORK DONE IN PHASE TWO

2.1 SYSTEM ARCHITECTURE DESIGN

System Design is a solution, how to approach to creation of a new system. This important phase is composed of several steps. It provides the understanding and procedural details for implementing the system recommended infeasibility study. Stress in on translating performance requirement into design specification design goes through logical physical stages of development. Logical design reviews the present physical, prepare input and output specification. These steps are as follow:

i. Problem definition.
ii. Input output specification.
iii. Data based designed.
iv. Modular program design.
v. Preparation of source code.

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

vi. Testing and debug.

INPUT DESIGN

Input Design plays a vital role in the life cycle of software development, it requires very careful attention of developers. The input design is to feed data to the application as accurate as possible. So inputs are supposed to be designed effectively so that the errors occurring while feeding are minimized. According to Software Engineering Concepts, the input forms or screens are designed to provide to have a validation control over the input limit, range and other related validations.This system has input screens in almost all the modules. Error messages are developed to alert the user whenever he commits some mistakes and guides him in the right way so that invalid entries are not made. Let us see deeply about this under module design.Input design is the process of converting the user created input into a computer-based format.

The goal of the input design is to make the data entry logical and free from errors. The error is in the input are controlled by the input design. The application has been developed in user-friendly manner. The forms have been designed in such a way during the processing the cursor is placed in the position where must be entered. The user is also provided with in an option to select an appropriate input from various alternatives related to the field in certain cases.Validations are required for each data entered. Whenever a user enters an erroneous data, error message is displayed and the user can move on to the subsequent pages after completing all the entries in the current page.

OUTPUT DESIGN

The Output from the computer is required to mainly create an efficient method of communication within the company primarily among the project leader and his team

members, in other words, the administrator and the clients. The output of VPN is the system which allows the project leader to manage his clients in terms of creating new clients and assigning new projects to them, maintaining a record of the project validity and providing folder level access to each client on the user side depending on the projects allotted to him. A new user may be created by the administrator himself or a user can himself register as a new user but the task of assigning projects and validating a new user rests with the administrator only.The application starts running when it is executed for the first time. The server has to be started and then the internet explorer in used as the browser. The project will run on the local area network so the server machine will serve as the administrator while the other connected systems can act as the clients. The developed system is highly user friendly and can be easily understood by anyone using it even for the first time.

Choosing the right output method for each user is another objective in designing output. Much output now appears on display screens, and users have the option of printing it out with their own printer. The analyst needs to recognize the trade-offs involved in choosing an output method. Costs differ; for the user, there are also differences in the accessibility, flexibility, durability, distribution, storage and retrieval possibilities, transportability, and overall impact of the data.

One of the most common complaints of users is that they do not receive information in time to make necessary decisions. Although timing isn't everything, it does play a large part in how useful output will be. Many reports are required on a daily basis, some only monthly, others annually, and others only by exception.

Designing Output to Serve the Intended Purpose

Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

All output should have a purpose. During the information requirements determination phase of analysis, the systems analyst finds out what user and organizational purposes exist. Output is then designed based on those purposes. You will have numerous opportunities to supply output simply because the application permits you to do so. Remember the rule of purposiveness, however. If the output is not functional, it should not be created, because there are costs of time and materials associated with all output from the system.

Designing Output to Fit the User

With a large information system serving many users for many different purposes, it is often difficult to personalize output. On the basis of interviews, observations, cost considerations, and perhaps prototypes, it will be possible to design output that addresses what many, if not all, users need and prefer. Generally speaking, it is more practical to create user-specific or user-customizable output when designing for a decision support system or other highly interactive applications such as those using the Web as a platform. It is still possible, however, to design output to fit a user's tasks and function in the organization, which leads us to the next objective.

Delivering the Appropriate Quantity of Output

Part of the task of designing output is deciding what quantity of output is correct for users. A useful heuristic is that the system must provide what each person needs to complete his or her work. This answer is still far from a total solution, because it may be appropriate to display a subset of that information at first and then provide a way for the user to access additional information easily.

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

The problem of information overload is so prevalent that it is a cliché, but it remains a valid concern. No one is served if excess information is given only to flaunt the capabilities of the system. Always keep the decision makers in mind. Often they will not need great amounts of output, especially if there is an easy way to access more via a hyperlink or drill-down capability.

2.2 SYSTEM REQUIREMENTS

2.2.1 HARDWARE REQUIREMENTS

Processor	: Pentium dual core
RAM	: 1 GB
Hard Disk Drive	: 80 GB
Monitor	: 17” Color Monitor
Key Board	: 108 Keys
Mouse	: 3 Buttons

2.2.2 SOFTWARE REQUIREMENTS

Front End/GUI Tool	: Microsoft Visual studio 2012
Operating System	: Windows Family
Language	: C#, ASP.NET
Application	: Web Application
Back End	: SQL Server 2005

CHAPTER 3

SYSTEM ORGANIZATION

3.1 USE CASE DIAGRAM

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

The use case diagram that was finally designed for the Student Admissions system can be seen in Figure 4.3.2. The most important step to fully understanding the requirements and scope of this system was the creation and subsequent refinement of this use case diagram. Many changes were made to the initial diagram before a consensus was met on the final design used.

3.2 COLLABORATION DIAGRAM

A collaboration diagram, also called a communication diagram or interaction diagram, is an illustration of the relationships and interactions among software objects in the Unified Modeling Language (UML). The concept is more than a decade old although it has been refined as modeling paradigms have evolved

` A collaboration diagram resembles a flowchart that portrays the roles, functionality and behavior of individual objects as well as the overall operation of the system in real time. Objects are shown as rectangles with naming labels inside. These labels are preceded by colons and may be underlined. The relationships between the objects are shown as lines connecting the rectangles. The messages between objects are shown as arrows connecting the relevant rectangles along with labels that define the message sequencing.

3.3 SEQUENCE DIAGRAM

Sequence diagrams are used to help quickly picture how the objects in a use case interact during the sequence of events in a use case . They do this by showing the behavior of the objects in the use case and the messages they pass. The main strength of sequence diagrams is the clarity with which they show what objects are making what calls and to whom, and which objects are doing what processing .

3.4 STATE CHART DIAGRAM

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

State chart diagram is one of the five UML diagrams used to model the dynamic nature of a system. They define different states of an object during its lifetime and these states are changed by events. State chart diagrams are useful to model the reactive systems. Reactive systems can be defined as a system that responds to external or internal events. State chart diagram describes the flow of control from one state to another state. States are defined as a condition in which an object exists and it changes when some event is triggered. The most important purpose of State chart diagram is to model lifetime of an object from creation to termination. State chart diagrams are also used for forward and reverse engineering of a system. However, the main purpose is to model the reactive system.

CHAPTER 4

TECHNOLOGIES USED

4.1 DOT NET OVERVIEW

Visual basic is the favorite programming environment of many users. When visual basic was originated, it created a revolution in Windows programming, and that revolution appears to this day. Windows programming was never as easy as it is using visual basic. It is so easy that programs are made right before the eyes and are executed. Visual Basic was derived from BASIC, and is an event-driven programming language. Programming in visual basic is done visually, which means that as you design, you will know how your application will look on execution. We can therefore change and experiment with design to meet your requirement.

Visual Basic 2015 has many powerful features that's are required in today's programming environment Visual basic introduced unheard-of ease to windows

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

programming and changed programming from a chore to something very fun. In time, Visual basic has gotten more complex as well as more powerful. Visual basic is used in a task-oriented way, which is the best way to write about programming.

PROJECT EXPLORER

This is the window that allows coordinating the parts of the program into folders for easy manipulations as all the parts of that project appear in a review. To pick the required part of the project, find that from the project explorer and double click on it, you will get that. Items such as forms can be added or removed from in the project explorer. The buttons at the top allows the user to switch between the different views. The left button displays the object code window, the middle button displays the object itself and the right button toggles the folder open and closed in the project explorer. So, project explorer gives us a valuable overview of our entire project, which is very useful when a project gets too large and contains many components.

PROPERTIES WINDOW

This is the window where the properties of an object can be set. When any object is selected in visual basic using mouse, its properties are displayed in the properties window. Properties of an object can be changed in two ways- either at design time or runtime. Properties appear in the properties window are set at design time.

4.2 DOT NET PROGRAMMING FEATURES

Writing a graphical user interface program is much easier in Visual Basic. Visual Basic has a very friendly environment, which helps to create forms, add contacts to the form and write code behind the form, very quickly and easily.

Visual basic also has an online help system. This Online help contains thousands of programming.

4.3 DOT NET FOUNDATION CLASS

Everything in Visual Basic is event-driven. Unlike earlier programming languages like COBAL and Pascal in which control was with the application, in Visual Basic user is in the control of the application. Every time the user clicks a command buttons or presses the mouse, an event stream is generated and coded that has been written behind the event is executed. After the execution of the code the control again comes back to the user.

GUI / WINDOW ENVIRONMENT

An application developed in Visual Basic has the looks and feel of windows application development system. It behaves like any other Windows program to which any user is accustomed, so, for any user it is not new and he feels very comfortable while using an application developed in Visual Basic.

INTERNET BASED APPLICATION

Visual Basic can be used to write Internet programs. Visual basic can be extended to other application by the use of Active X controls, Dynamically Linked Libraries (DLL's), and Add-Ins.

4.4 OVERVIEW OF DOT NET PROGRAMMING

Microsoft Visual Studio is an integrated development environment (IDE) from Microsoft. It is used to develop computer programs for Microsoft Windows, as well as web sites, web apps, web services and mobile apps. Visual Studio uses Microsoft software development platforms such as Windows API, Windows Forms, Windows Presentation

Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

Foundation, Windows Store and Microsoft Silverlight. It can produce both native code and managed code.

Visual Studio includes a code editor supporting IntelliSense (the code completion component) as well as code refactoring. The integrated debugger works both as a source-level debugger and a machine-level debugger. Other built-in tools include a code profiler, forms designer for building GUI applications, web designer, class designer, and database schema designer. It accepts plug-ins that enhance the functionality at almost every level—including adding support for source control systems (like Subversion) and adding new toolsets like editors and visual designers for domain-specific languages or toolsets for other aspects of the software development lifecycle (like the Team Foundation Server client: Team Explorer).

CHAPTER 5

IMPLEMENTATION AND RESULTS

5.1 IMPLEMENTATION

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective. The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

The following chapter looks at how the system implementation progressed over the course of the build. It starts off by detailing the plan that was used to order how the

implementation would progress. The chapter then covers some of the important aspects of implementation that were carried out throughout the building of the system. The data-tier and then the web tier will be looked at in turn, and for the web-tier the implementation of each aspect of the Model-View-Controller will be considered. Finally, the testing that was carried out on the system will be examined and the effect the results had on the system will be looked.

Designing Plan

As was stated in the Technology and Literature Survey chapter the methodology being followed for the development of this system is the Dynamic Systems Development Method (DSDM). Many of the core principles of the DSDM come about from the idea that the designing and building of a system should be an incremental process. To benefit most from the methodology being used it was decided that the implementation phase of the development would be carried out in an incremental fashion and to do this the system would have to be split into distinct sections. Due to the manner in which the requirements were gathered it was very easy to view the system in such a way. It was decided that the best way to split the system up would be by the main tasks that the system was required to achieve and it was hoped that by doing so each increment of the build would be kept to a manageable level.

The testing for the system will also be carried out as it is built with a slightly more thorough test at the end of each cycle. It is hoped that by doing this the system will nearly always be in a relatively stable state, and that if when it gets to the deadline not all the functionality is present this would not cause major problems with the final implementation.

Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

5.1.1 Data Flow diagram

The DFD was first developed by Larry Constantine as a way of expressing system requirement in a graphical form. A DFD also known as bubble chart has a purpose of clarifying system requirement and identifying major transformation that will become the program in the system design.The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of the input data to the system, various processing carried out on these data, and the output data is generated by the system.

- The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.
- DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.
- DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.

5.2 MODULES

5.2.1 Multi-Objective Optimization

5.2.2 Assignment of Documents to Different Clusters

5.2.3 Use Genetic Operators

5.2.4Application of Semi-supervision on NSGA-II Algorithm

MODULES DESCRIPTIONS

5.2.1 Multi-Objective Optimization

Simultaneously optimizing several objective functions is known as multi-Objective optimization (MOO) (Deb, 2001). In general the objective functions used in MOO are conflicting in nature. A real-life ex-ample could be buying a car where the objectives are : i) minimizing cost and ii) maximizing comfort. In mathematical terms, a MOO problem can be formally stated as: Finding the vectors of decision vari-ables $x = [x_1, x_2, x_3, ..., x_n]^T$ which will satisfy m inequality constraints: $g_i(x) \geq 0$, i = 1, 2, ...m and p equality constraints $h_j(x) = 0$, j = 1, 2, ..., p and simultaneously optimize M objective functions $f_1(x)$, $f_2(x)$, ..., $f_M(x)$.

5.2.2 Assignment of Documents to Different Clusters

In our experiments we have used cosine and Euclidean distance as separate parameters to assign the documents in respective clusters. For each document we determine any of the available distance measures with respect to all the cluster medoids (encoded in a particular chromosome). Finally the document is assigned to that cluster medoid (m_i) with respect to which it is having the minimum distance. Once the assignment has been done for all the documents, the new cluster medoids are calculated based on the new clusters formed. These new medoids replace the existing medoids represented in that particular chromosome.= $\text{argmin}^{K}_{j=1}\, d(x, m_j)$.

Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

5.2.3 Use Genetic Operators

We use classical mutation and crossover operators as proposed in NSGA-II (Deb et al., 2013) to bring diversity in our population. Suppose there is a chromosome (2 4 5 7 8 9) representing a parent chromo-some. In a mutation two documents are selected and exchanged.(2 4 5 7 8 9) => (2 9 5 7 8 4)In the case of crossover operation the bits are exchanged between parent chromosomes to produce off-springs. Once a crossover point is selected, the permutation till this point is copied from the first parent, then the second parent is scanned and, if the number is not yet in the offspring, it is added.

5.2.4 Application of Semi-supervision on NSGA-II Algorithm

Here we perform some modifications in the selection step of NSGA-II to take care of the available supervised information in terms of must-link and cannot-link constraints. The computation of non-dominated fronts depends not only on the objective functions (XB and I indices) but on the available constraints (must-link, cannot-link) also. A must-link constraint ensures that two instances should remain in the same cluster as a cannot-link constraint ensures that two instances should be in two different clusters. It is assumed that the documents lying in the same cluster obey must-link and the different clusters obey.

CHAPTER 6

CONCLUSION & FUTURE WORK

In this paper, we propose a new semi-supervised clustering ensemble approach, which is referred toas the incremental semi-supervised clustering ensemble approach

(ISSCE). Our major contribution is the development of an incremental ensemble member selection process based on a global objective function and a local objective function.In order to design a good local objective function, we also propose a new similarity function to quantify the extent to which two sets of attributes in the subspaces are similar to each other.We conduct experiments on 6 real-world datasets from the UCI machine learning repository and 12 real-world datasets of cancer gene.

REFERENCES

[1] A. Strehl, J. Ghosh, "Cluster Ensembles — A Knowledge Reuse Framework for Combining Multiple Partitions", Journal of Machine Learning Research, vol. 3, pp. 583-617, 2002.

[2] L.I. Kuncheva, and Dmitry Vetrov, "Evaluation of Stability of k-Means Cluster Ensembles with Respect to Random Initialization", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 11, pp. 1798-1808, 2006.

[3] A.P. Topchy, A.K. Jain, and W.F. Punch, "cluster ensembles: Models of Consensus and Weak Partitions", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 12, pp. 1866-1881, 2005.

[4] H.G. Ayad, M.S. Kamel, "Cumulative Voting Consensus Method for Partitions with Variable Number of Clusters", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 30, Issue 1, pp.16–173, Jan. 2008.

[5] N. Iam-On, T. Boongoen, S. Garrett, C. Price, "A Link-Based Approach to the Cluster Ensemble Problem", IEEE Transactions on Pattern Analysis and Machine Intelligence , vol. 33, no. 12, pp. 2396-2409, 2011.

Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

[6] N. Iam-On, T. Boongoen, S. Garrett, C. Price, "A Link-Based cluster ensemble approach for categorical data clustering", IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 3, pp. 413-425, 2012.

[7] Y. Yang, K. Chen, "Temporal Data Clustering via Weighted Clustering Ensemble with Different Representations", IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 2, pp. 307-320, 2011.

[8] Zhiwen Yu, Hantao Chen, Jane You, Guoqiang Han, Le Li, "Hybrid Fuzzy Cluster Ensemble Framework for Tumor Clustering from Biomolecular Data", IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 10, no. 3, pp. 657-670, 2013.

[9] Zhiwen Yu, Le Li, Jane You, Guoqiang Han, "SC3: Triple spectral clustering based consensus clustering framework for class discovery from cancer gene expression profiles", IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol.9, no.6, pp.1751-1765, 2012.

[10] Zhiwen Yu, Hantao Chen, Jane You, Hau-San Wong, Jiming Liu, Guoqiang Han, Le Li, "Adaptive Fuzzy Consensus Clustering Framework for Clustering Analysis of Cancer Data", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2015 (DOI:10.1109/TCBB.2014.2359433)