DISCRETIZING THE PREPROCESSED AUTOMATED BLOOD CELL COUNTER DATA USING CHI MERGE ALGORITHM IN CLINICAL PATHOLOGY

D.Minnie[1] and S.Srinivasan[2]

[1]Department of Computer Science, Madras Christian College, Chennai, India
minniearul@yahoo.com
[2]Department of Computer Science and Engineering, Anna University Regional Office, Madurai, India
sriniss@yahoo.com

ABSTRACT

This paper applies the preprocessing phases of the Knowledge Discovery in Data-bases to the automated blood cell counter data and creates discrete ranges of blood cell counter data that can be used in grouping data using classification, clustering and association rule generation. The functions of an automated blood cell counter from a clinical pathology laboratory and the phases in Knowledge Discovery in Databases are explained briefly. Twelve thousand records are taken from a clinical laboratory for processing. The preprocessing steps of the KDD process are applied on the blood cell counter data. This paper applies the Chi Merge algorithm on the blood cell counter data and generates discretized data representing ranges of values for the data.

KEYWORDS

Clinical Pathology, Blood Cell Counter, Knowledge Discovery in Databases, Data Mining, Discretization, Chi Merge algorithm

1. INTRODUCTION

Clinical Pathology is associated with monitoring diseases of patients by conduct-ing tests on various body fluids. The fluids are either tested using manual procedure or an automated procedure. A Blood Cell Counter is an automated system that generates blood test results. The data contains noise such as missing values and the data is to be cleaned. The preprocessing phase of the Knowledge Discovery in Databases (KDD) Techniques is applied on the blood cell counter data to prepare the Blood Cell Counter Medical Data for efficient data mining. KDD [1], [2] is used to generate meaningful results from data and hence it is applied on medical data to generate knowledge.

2. AUTOMATED BLOOD CELL COUNTER DATA

A Blood Cell Counter [3] is an automated machine that can be loaded with dozens of blood samples at a time and the Complete Blood Count (CBC) or Full Blood Count (FBC) of the given blood samples are generated as a report. The number of red blood cells, white blood cells and platelets are some of the blood counts generated. The results are either printed directly or are stored in the computer for later use.

The 12,000 cell counter data are collected from a Clinical Pathology department of a reputed hospital. The data is present as an excel file and the data is used to generate association rules among the various attributes of the ABCC Database.

DOI : 10.5121/ijitcs.2012.2504

2.1. Automated Blood Cell Counter Data Format

The Blood Cell Counter Data is an excel file. The Blood Cell Counter generates files as output and the files consist of values for various attributes for each sample of blood.

The attributes of the records considered for further processing include PId, SId, PName, PAge, PGender, RDate, RTime, Hg count, MCH, MCHC, MCV, MPV, PCT and RDW.

2.2. Sample Blood Cell Counter Data

A sample of the Automated Blood Cell Counter in the Excel format is shown in Fig. 1.

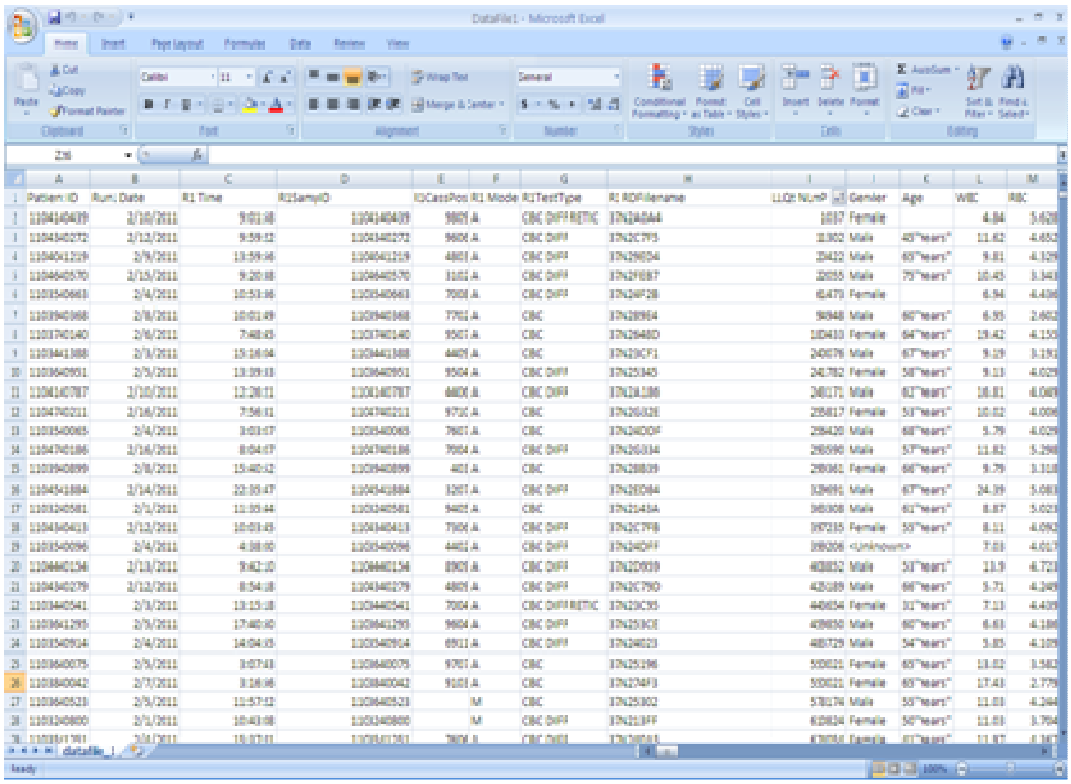


Figure 1. Sample Automated Blood Cell Counter Data

2.3. Blood Cell Counter Data Attributes

The Automated Blood Cell Counter generates values for various attributes such as RBC, WBC counts and so on. A few of the attributes are selected for processing and are shown in table 1.

Table 1. Automated Blood Cell Counter Attributes.

Attribute Name	Attribute Description
PID	Patient Id
PNAME	Patient Name
PAGE	Patient Age
PGENDER	Patient Gender
SID	Sample Id
RD	Recorded Date

RT	Recorded Time
Hg	Hemoglobin Count
MCH	Mean Corpuscular Haemoglobin
MCHC	Mean Corpuscular Haemoglobin Concentration
MCV	Mean Corpuscular Volume
MPV	Mean Platelet Volume
PCT	Prothrombin Consumption Time
RDW	Red cell Distribution Width

3. KNOWLEDGE DISCOVERY IN DATABASES (KDD)

KDD consists of the processes Data Cleaning, Data Integration, Data Selection, Data Transformation, Data Mining, Generation of Patterns and Knowledge Interpretation for effective knowledge generation and is shown in Fig.2.

In Data Cleaning the irrelevant data are removed from the collected data. In Data Integration multiple sources may be combined into a common source. The Data Selection process is involved with the selection of data relevant to the analysis and extracting them from the integrated data. The selected data is transformed to the appropriate form for the mining procedure.

The data is divided into various ranges of values using discretization techniques. The discretized data is used to generate concept hierarchies which are helpful in the decision tree technique of classification. Chi Merge is an interesting algorithm used to discretize given data in to various ranges.

The process of extracting useful and implicit information from the transformed data is referred to as Data Mining. In Pattern Evaluation interesting patterns are identified from the processed data. The discovered knowledge is visually represented to the user in the Knowledge Representation process.

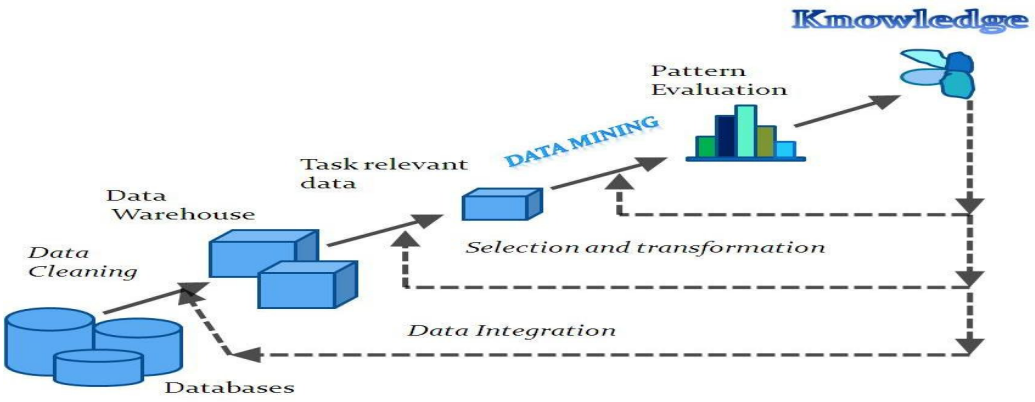


Figure 2. Knowledge Discovery in Databases (KDD)

Data Mining is the Knowledge Discovery phase of KDD and it is the process of extracting implicit, useful, previously unknown, non-trivial information from data [1]. The techniques

involved in Data Mining are grouped as Classification, Clustering, Association Rules and Sequences.

Classification is a supervised learning process and it maps data into known classes using Decision Trees, Neural Networks and Genetic Algorithms. Clustering is an unsupervised learning and it groups similar data into unknown clusters using K-Means, Nearest Neighbour and various other algorithms. Association Rule Mining (ARM) [4] uncovers relationships among data in a database.

Association Rule Mining (ARM) is used to find frequent patterns, associations and correlations among sets of items in databases and any other information repositories. Association Rule correlates the presence of one set of items with that of another set of items in the same transaction. The quality of an Association Rule is measured using its support and confidence values and several efficient methods are developed [5] to generate association rules.

4. RELATED WORK

Quality control is used in all laboratories to check errors and it plays a vital role in Clinical Pathology. The role of auto verification of results [6] in a laboratory in-formation system is very important as the normal results can be generated at the speed of an automated machine. The abnormal results have to be analysed using various techniques. Specimen mislabelling is one of the errors present in Transfusion Medicine and it can be reduced by collecting and trending the data on mislabelled samples with timely feedback to patient care [7].

Various combinations of Data Mining classification algorithms are used on medical data for efficient classification of data [8]. [9] presents ways of using sequences of clustering algorithms to mine temporal data. Association Rule Mining is used to diagnose diseases [10], [11] and risk patterns [12] from medical data. Taxonomy is used in certain cases to establish associations between different items in a data base [13]. Apriori algorithm is used to find frequent item sets in a database and to generate Association Rules from the frequent item sets [14]. A survey of various Data Mining Tools is presented in [15] and each of the tools is designed to handle a specific type of data and to perform a specific type of task.

Medical data is taken most of the times from medical records [16] and the data is found to be heterogeneous [17] in nature. The privacy issues [17] are to be finalized before handling medical data. The data that is taken from the Blood Cell Counter for our work is De-identified and the patient id and names are changed by the Clinical Pathology department before supplying the medical data for analysis.

5. METHODOLOGY

The methodology applied for this research work is summarized as follows:

- Analyze Automated Blood Cell Counter functionality
- Collect Blood Cell Counter Data
- Apply KDD Data Cleaning Process
- Apply KDD Data Transformation Process
- Apply Data Reduction Process
- Apply Chi-Merge Data Discretization Process

5.1. Analyse Automated Blood Cell Counter Functionality

The functionality of the Automated Blood Cell Counter are briefly explained in Chapter 2 of this paper.

5.2. Automated Blood Cell Counter Data Collection

Twelve thousand cell counter data are collected from a Clinical Pathology department of a reputed hospital. The data is present as an excel file and the data is used to generate association rules among the various attributes of the ABCC Database

5.3. Apply KDD Data Cleaning Process

The process of detecting and correcting or removing corrupt or inaccurate records from a record set, table, or database is Data Cleaning. The Blood Cell Counter data contained missing fields and such records were not required for further analysis.

5.4. Apply KDD Transformation Process

The process of converting the data in one form in to another form is data transformation. It involves normalization and aggregation processes.

5.5. Apply Data Reduction Process

The process of obtaining reduced representation in volume but that produces the same analytical result is data reduction. Various strategies such as data cube aggregation, dimensionality reduction, data compression, numerosity reduction and discretization can be applied on data to reduce the volume of data.

5.6. Apply Chi-Merge Data Discretization Process

Chi Merge is an algorithm used to discretize data and it uses Chi Square statistics. It is applied on the attributes of a database. The adjacent pairs of values are compared to find the similarity among the data using chi square test. If the data are similar they are kept in same interval and they are put in different intervals if the similarity between the two is very less. In other words they are not similar.

The Chi Merge Discretization process is explained in the following steps:

- Sort and order the attributes that are to be grouped. For example sort the database using the attribute age or the RBC count.
- Find the Threshold for merging.
- Initially each record is assigned a unique interval. That interval should accommodate the value of the attribute of that record.
- Calculate the Chi Square test on each interval.
- Find the Chi Square value of all intervals that are less than the Threshold value.
- Find the intervals with the smallest Chi Square value
- Merge the adjacent intervals corresponding to the lowest Chi Square value.
- Repeat the tests until there are no more intervals that can satisfy the Chi Square test.

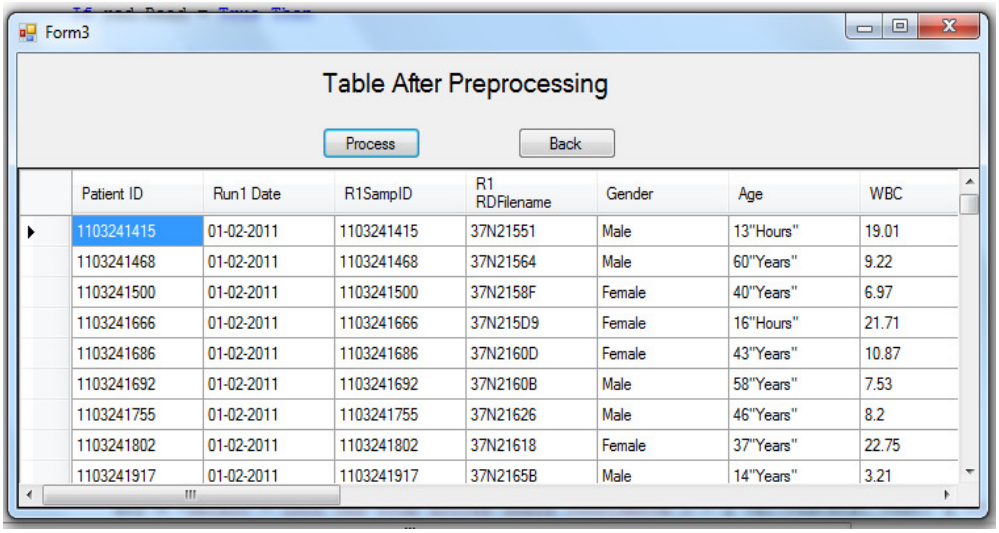
6. RESULTS AND DISCUSSIONS

The collected Automated Blood Cell Counter Data is subjected to the Data Cleaning, Data Selection, Data Transformation and Data Reduction processes. Knowledge such as Association

Rules, Clusters and Classifications are generated by applying Data Mining algorithms to the Automated Blood Cell Counter Data.

6.1. Data Cleaning

The attributes RDate, RTime, Hg count, MCH, MCHC, MCV, MPV, PCT and RDW were required for further processing and hence the records without these fields were removed. The resultant excel file contained the records with patient id, gender, age, date and time of results and the blood count fields were selected for further processing and is given in fig.3.



Patient ID	Run1 Date	R1SampID	R1 RDFilename	Gender	Age	WBC
1103241415	01-02-2011	1103241415	37N21551	Male	13"Hours"	19.01
1103241468	01-02-2011	1103241468	37N21564	Male	60"Years"	9.22
1103241500	01-02-2011	1103241500	37N2158F	Female	40"Years"	6.97
1103241666	01-02-2011	1103241666	37N215D9	Female	16"Hours"	21.71
1103241686	01-02-2011	1103241686	37N2160D	Female	43"Years"	10.87
1103241692	01-02-2011	1103241692	37N2160B	Male	58"Years"	7.53
1103241755	01-02-2011	1103241755	37N21626	Male	46"Years"	8.2
1103241802	01-02-2011	1103241802	37N21618	Female	37"Years"	22.75
1103241917	01-02-2011	1103241917	37N2165B	Male	14"Years"	3.21

Fig. 3. Cleaned data

6.2. Data Selection

The attribute RBC count is selected for discretizing the data. The continuous numeric value of the RBC count is replaced with the range of values identified in the Data Discretization process.

6.3. Data Transformation

The Data Transformation stage is involved with converting data from a source data format into destination data format. The data in excel format is converted into sql format for efficient processing of the data.

6.4. Data Reduction using Chi-Merge Algorithm

The Chi Merge algorithm is applied on the RBC count of the blood cell counter data and the generated ranges of values are shown in Fig.4

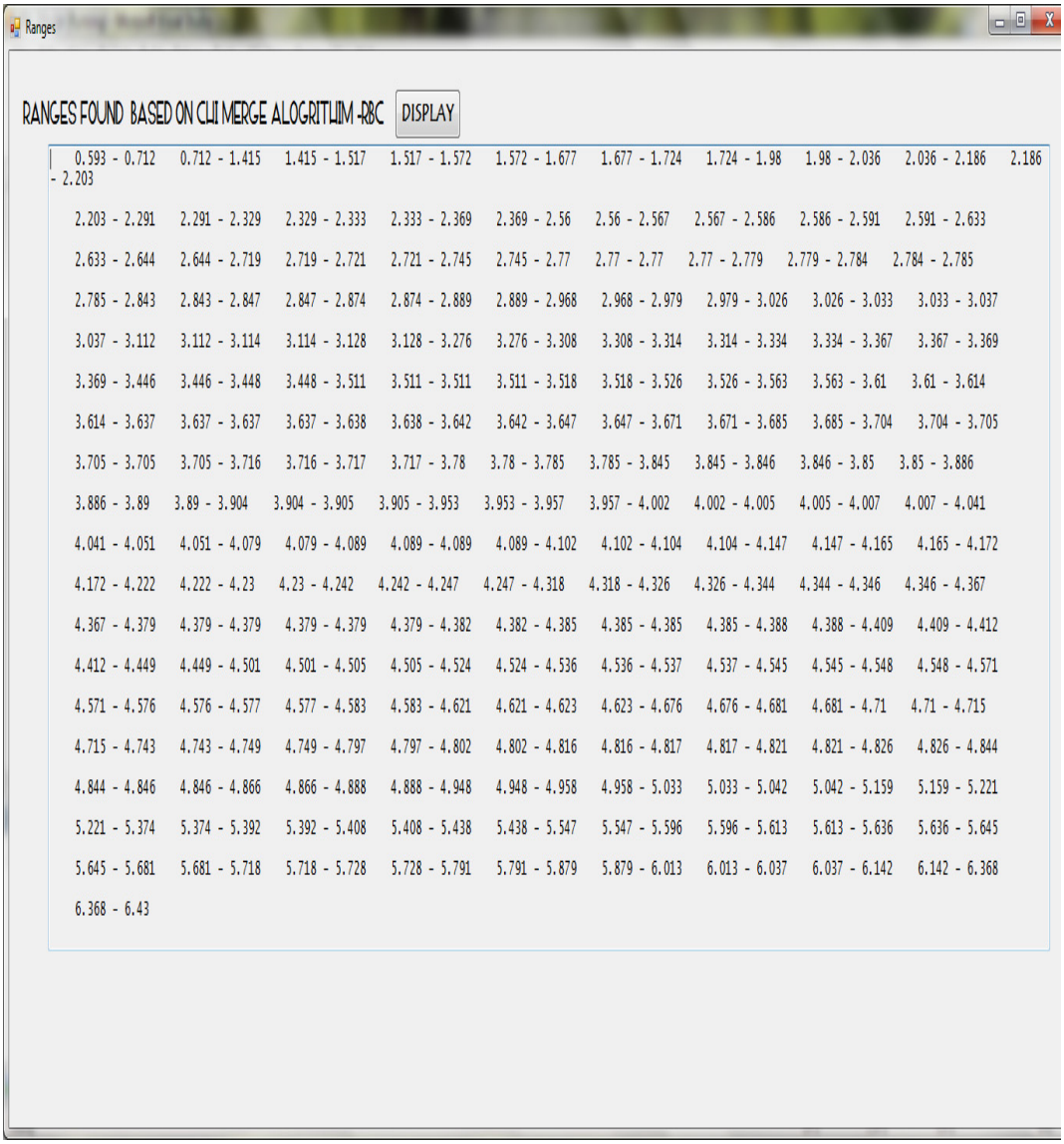


Fig.4. Discretized Data

7. CONCLUSIONS

A brief study of Blood Cell Counter and Blood Cell Counter data is presented in the paper. The blood cell counter data was analyzed and few attributes were se-lected for processing, based on the knowledge given by the Clinical Pathologist. The KDD steps namely Data Cleaning, Integration, Selection, Transformation, and Mining were explained and were applied on the Blood Cell Counter Data to convert the raw data into a transformed data that was used for generating knowledge from the system. The data is discretized using Chi Merge algorithm.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Joy John Mammen, MD, Department of Transfusion Medicine and Immunohematology, Christian Medical College, Vellore, Tamilnadu, India for

sharing his knowledge in Clinical Pathology, especially the functions of the Blood Cell Counter and also for providing the De identified blood cell counter data.

REFERENCES

[1] Jaiwei Han, Michelinne Kamber , Data Mining : Concepts and Techniques, Morgan Kaufmann Publishers, Second Edition, 2006

[1] Margaret H.Dunham, Data Mining: Introductory and Advanced Topics, Pearson Education, 2007.

[2] Automated Blood Cell Counter: www.medscape.com

[3] Dion H.Goh and Rebecca P.Ang. An Introduction to Association rule mining: An application in counseling and help-seeking behavior of adolescents. Behaviour Research Methods, 39(2), 2007, pp. 259-266.

[4] Rakesh Agrawal, T. Imielinski, A. Swami, Mining Associations between Sets of Items in Large Databases, Proc. of the ACM-SIGMOD 1993 Int'l Conference on Management of Data, Washington D.C., May 1993, pp. 207 - 216.

[5] Dale J. Duca, Auto Verification in a Laboratory Information System, Laboratory Medicine, January 2002, number 1, Volume 33, pp. 21 – 25.

[6] Karen Quillen and Kate Murphy, Quality Improvement to Decrease Specimen Mislabeling in Transfusion Medicine, Archives of Pathology and Laboratory Medicine, Vol 130, August 2006, pp. 1196 - 1198.

[7] Alp Aslandogan Y. and Gauri A.Mahajani, Evidence Combination in Medical Data Mining, Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04), Volume 2, 2004, pp. 465 – 469

[8] Rakesh Agrawal, T. Imielinski, A. Swami, Database Mining: A Performance Perspective, IEEE Transactions on Knowledge and Data Engineering, Volume 5 Issue 6, December 1993, pp. 914 – 925.

[9] Massoud Toussi, Jean-Baptiste Lamy, Philippe Le Toumelin, and Alain Venot, Using data mining techniques to explore physicians' therapeutic decisions when clinical guidelines do not provide recommendations: methods and example for type 2 diabetes, BMC Medical Informatics and Decision Making 2009; pp. 9:28

[10] Sengul Dogan and Ibrahim Turkoglu. Diagnosing Hyperlipidemia using Association rules, Mathematical and Computational Applications, Association for Scientific Research, Vol.13,No. 3, 2008, pp. 193-202

[11] Jiuyong Li, Ada Wai-chee Fu and Hongxing He Et. Al, Mining risk Patterns in Medical data, KDD'05, Chicago, Illinois, USA, 2005, pp. 770 – 775.

[12] Ramakrishnan Srikant and Rakesh Agrawal, Mining Generalized Association Rules, Proceedings of the 21st International Conference on Very Large Data Bases, Zurich, Swizerland, September 1995

[13] Rakesh Agrawal and Ramakrishnan Srikant, Fast Algorithms for Mining Association Rules, Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile, September 1994

[14] Michael Goebel, Le Gruenwald, A Survey of Data Mining and Knowledge Discovery Software Tools, SIGKDD Explorations, ACM SIGKDD, June 1999.

[15] Patricia Cerrito, John C. Cerrito, Data and Text Mining the Electronic Medical Record to Improve Care and to Lower Costs, Proceedings of SUGI 31, March 26 – 29, 2006 paper 077-31, 2006

[16] Cios KJ, Moore GW, Uniqueness of Medical Data Mining, Artificial Intelligence in Medicine, 2002 Sep-Oct; 26(1-2): 2002, pp. 1- 24.

Authors

Ms. D. Minnie, M.C.A., (Ph.D in Computer Science – registered in 2009),

Head In-Charge, Department of Computer Science, Madras Christian College, Chennai, India.

Presented papers in National and International Conferences.

Chairperson/Member in various Academic boards such as Board of Studies, Academic Council, Board of Examiners, Expert Committees.

2 decades of teaching experience and 6 years of research experience.

Dr. S. Srinivasan, M.Sc.(Maths), M.Tech (CSE), Ph.D.(CSE),

Professor and Head, Department of Computer Science and Engineering, Anna University Regional Office, Madurai, India.

No of Ph.D. students supervised/registered: 8.

Presented papers in National and International Conferences.

Chairperson/Member in various Academic boards such as Board of Studies, Academic Council, Board of Examiners, Expert Committees.

2 decades of teaching and research experience.

Membership in Professional bodies: ISTE, CSI