Volume 6, Issue 3, March 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Multi-label Data Categorization using Ant Colony Optimization and Relevance Clustering Technique

Ashwini Digambar, Prof. (Dr.) L. K. Vishwsmitra
CSE, Patel College of Engineering, Bhopal, Rajiv Gandhi Proudyogiki Vishwavidyalaya,
Bhopal, India

Abstract— ***A text classification assignment consists of the training phase and the text categorization phase. The former includes the feature extraction procedure and the indexing process. Due to the existence of relevance and co-occurrence among labels in multi-label classification, single-label classification methods cannot be used to directly address the multi-label classification problem. A large body of research has been carried out to explore effective and efficient multi-label classification approaches which are generally grouped into two main categories: problem transformation methods and algorithm adaptation methods. However, most of these methods neglect a fact that there exists some uncertainty during the process of classification.***

Keywords— Data mining, computational geometry, data mining, information retrieval, string matching

I. INTRODUCTION

Conventional classification approaches assume that each instance is associated with only one class label within a number of candidate classes. However, many real-world applications often involve the scenario where each instance can be assigned with a set of multiple labels. For example, in image annotation, one image can be tagged with a set of multiple words, such as urban, building, and road, indicating the contents of the image. Nowadays, we notice that multi-label classification methods are increasingly required by modern applications, such as protein function classification, music categorization and semantic scene classification. In semantic scene classification, a photograph can belong to more than one conceptual class at the same time, such as sunsets and beaches. Similarly, in music categorization, a song may belong to more than one genre.

II. DATA MINING

Data Mining is defined as mining of knowledge from huge amount of data. Using Data mining we can predict the nature and behaviour of any kind of data. The past two decades has seen a dramatic increase in the amount of information being stored in the electronic format. This accumulation of data has taken place at an explosive rate. It was recognized that information is at the heart of the business operations and that decision makers could make the use of data stored to gain the valuable insight into the business. DBMS gave access to the data stored but this was only small part of what could be gained from the data. Analysing data can further provide the knowledge about the business by going beyond the data explicitly stored to derive knowledge about the business. Learning valuable information from the data made clustering techniques widely applied to the areas of artificial intelligence, customer – relationship management, data compression, data mining, image processing, machine learning, pattern recognition, market analysis, and fraud – detection and so on. Cluster Analysis of a data is an important task in Knowledge Discovery and Data Mining. Clustering is the process to group the data on the basis of similarities and dissimilarities among the data elements. Clustering is the process of finding the group of objects such that object in one group will be similar to one another and different from the objects in the other group. A good clustering method will produce high quality clusters with high intra cluster distance similarity and low inter cluster distance similarity.

III. DATA MINING METHODS

There are several major data mining techniques have been developed and used in data mining:

- Association
- Classification
- Clustering
- Prediction
- Sequential Patterns

Text mining can be broadly defined as a knowledge-intensive process in which a user interacts with a document collection over time by using a suite of analysis tools. In a manner analogous to data mining, text mining seeks to extract useful information from data sources through the identification and exploration of interesting patterns. In the case of text mining, however, the data sources are document collections, and interesting patterns are found not among formalized database records but in the unstructured textual data in the documents in these collections. Clustering is the process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. It helps users to understand the natural grouping or structure in a dataset. A good clustering method will produce high quality clusters in which the intra-class (i.e., intra-clusters) similarity is high and the inter-class similarity is low. The quality of clustering result also depends on both the similarity measure used by the method and its implementation. The quality of a clustering method is also measured by its ability to discover some or the entire hidden pattern.

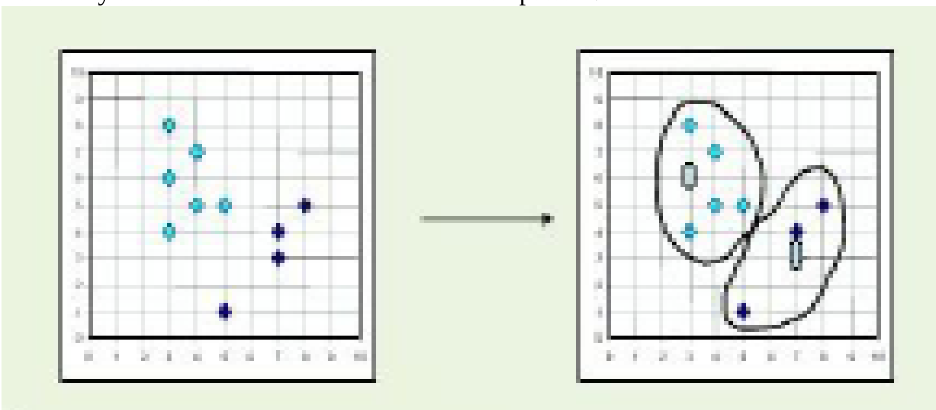

Figure 1.1: The result of cluster analysis.

Cluster analysis groups objects (observations, events) based on the information found in the data describing the objects or their relationships. The goal is that the objects in a group will be similar (or related) to one other and different from (or unrelated to) the objects in other groups. The greater the similarity (or homogeneity) within a group and the greater the difference between groups, the "better" or more distinct the clustering.

IV. LITERATURE SURVEY

Shie-Jue Lee, and Jung-Yi Jiang[1], We propose a fuzzy based method for multi label text classification in which a document can belong to one or more than one category. In text categorization, the number of the involved features is usually huge, causing the curse of the dimensionality problem. Besides, a category can be a non convex region, which is a union of several overlapping or disjoint sub regions. An automatic classification system, thus, may suffer from large memory requirements or poor performance.

Fran_cois Queyroi, Maylis Delest, Jean-Marc F_edou, Guy Melan_con[2], In our opinion, a theory supporting these multilevel clustering approaches is yet to be developed. Indeed, to the best of our knowledge there are no known optimization multilevel criteria guiding the construction of a hierarchy of clusters: the hierarchy basically is an fact of an iterative procedure. The main results of this paper contribute to such a multilevel clustering theory, by designing and studying a multilevel modularity measure for hierarchically clustered graphs, explicitly taking the nesting structure of clusters into account. The multilevel modularity we propose generalizes a modularity measure introduced by Man coridis et al. in the context of reverse software engineering. The measure we designed recursively traverses the hierarchy of clusters and computes a one variable polynomial encoding the intra and inter-cluster densities appearing at all levels in a hierarchical clustering. The resulting polynomial reacts how the graph combines with the hierarchy of clusters and can be used to assess the quality of a hierarchical clustering.

Min-Ling Zhang and Zhi-Hua Zhou[4], Multi-label learning studies the problem where each example is represented by a single instance while associated with a set of labels simultaneously. During the past decade, significant amount of progresses have been made towards this emerging machine learning paradigm. This paper aims to provide a timely review on this area, with emphasis on state-of-the-art multi-label learning algorithms. Firstly, fundamentals on multi-label learning including formal definition and evaluation metrics are given. Secondly and primarily, twelve representative multi-label learning algorithms are scrutinized under common notations, with corresponding analyses and discussions. Thirdly, several extended topics on multi-label learning are briefly summarized. As a conclusion, online resources and open research problems on multi-label learning are outlined for reference purposes.

V. PROPOSED METHODOLOGY AND ARCHITECTURE

In this we proposed a feature optimization and feature selection technique for multi-level classification technique. The multi-level classification technique suffered a problem of feature selection and feature optimization. The process of feature optimization reduces the unwanted and unused feature of data during the process of classification. For the optimization of feature used ant colony optimization technique. Ant colony optimization technique well knows optimization technique. The process of ant colony optimization technique followed the principle of biological ants. The artificial ants behave just like biological ants and find the best solution for the process of optimization. In the continuity of chapter discuss the feature optimization technique, clustering mapping, ant colony optimization and proposed algorithm and proposed model.

Cluster Mapping Technique

Multi-label categorization method processes multi-label data directly. Two phases, training and testing, are involved in the method. The training phase is divided into four steps: fuzzy transformation, fuzzy relevance clustering, cluster-to-category mapping, and finding thresholds. In the training phase, a classification model is created from a set of training data. High-dimensional training data are transformed to low-dimensional fuzzy relevance vectors. The fuzzy clusters are dynamically computed, and then linearly mapped to categories, and finally, a set of thresholds are computed. In the testing phase, the trained model obtained in the training phase is used to assign categories to unseen data. This phase is also divided into four steps: fuzzy transformation, calculating cluster memberships, cluster-to-category mapping, and applying thresholds. An unseen document, after fuzzy transformation, is applied to all clusters and a similarity is found to each cluster. Then, the calculated similarities are fed to the linear mapping to get a similarity vector to categories. Finally, the similarity vector to categories is compared with the thresholds to determine which of the categories are assigned to the document.

Our categorization method has several advantages. Since dimension reduction is performed by fuzzy transformation, the curse of the dimensionality problem is avoided. Furthermore, the region a category covers is a combination of several sub-regions. Therefore, complex category boundaries are possible. Other methods may have the limitation that they can only allow decision regions to be convex. Our proposed method overcomes this limitation.

- Fuzzy Transformation

Usually, m, which is the number of features in T , is large. This results in a sparse distribution of the training data in a high-dimensional space, which causes the curse of the dimensionality problem. By fuzzy transformation, a document with m dimensions is transformed to a fuzzy relevance vector with p dimensions, where p is the number of categories in C. Usually; p is much smaller than m. Therefore, dimension reduction is achieved and the curse of dimensionality problem is avoided.

- Fuzzy Relevance Clustering

Given n fuzzy relevance vectors x(1) , x(2) ,..., and x(n) , a clustering algorithm is applied to group these vectors into clusters. A number of clustering algorithms have been proposed [3]. We adopt the clustering scheme proposed in [6] with a modification. This algorithm is efficient. The participating vectors are scanned only once. Besides, the number of clusters need not be specified in advance by the user. Rather, new clusters are created automatically and incrementally.

- Cluster-to-Category Mapping

As mentioned earlier, a category may cover a region that consists of a number of sub regions. The clusters obtained by fuzzy relevance clustering are possible candidate sub regions since the constituent vectors in a cluster are closely related to each other. Therefore, it is reasonable to explore a relationship between clusters and categories. For simplicity, the relationship is assumed to be a linear mapping.

- Finding Thresholds

When the category similarity vector a1,a2 ,...,ap is obtained for a document, we would like to produce p output values, 1 , o2 ,...,op , for the document, such that oj =1 indicates that the document belongs to category cj , 1 ≤ j ≤ p. Several outputs can be 1, which indicate that the document belongs to several categories. Therefore, multi-label classification is achieved. For this, we have a hard-limiting function between each aj − oj pair, such that

$$O_j = \begin{cases} 1, if\, a_{j \geq \tau_j} \\ 0, otherwise \end{cases} \quad \text{.................................}(4.1)$$

Ant Colony Optimization

ACO makes probabilistic decision in terms of the artificial pheromone trails and the local heuristic information. This allows ACO to explore larger number of solutions than greedy heuristics [36, 37]. Another characteristic of the ACO algorithm is the pheromone trail evaporation, which is a process that leads to decreasing the pheromone trail intensity over time. Pheromone evaporation helps in avoiding rapid convergence of the algorithm towards a sub-optimal region. ACO algorithm is used for feature subset selection in that ACO makes probabilistic decisions in terms of the artificial pheromone trails and the local heuristic information. Feature subset is generated using ACO by eliminating redundant and noisy features result in good predictive performance and reduced computation [37]. OAO and OAA class based on SVM technique is efficient process, but this SVM based feature selection generate result on the unclassified of data. When the scale of data set increases the complexity of preprocessing is also increases, it is difficult to maintain correct matching classification. Ant Colony Optimization (ACO) meta-heuristic is an effective tool in finding quality data [32] and that’s the main reason to use it as a feature selection for SVM. ACO-based feature selection procedure eliminates insignificant features from the feature. Next, the reduced feature is passed to SVM procedure.

Feature Optimization

Text data feature extraction play an important role in multi-label text categorization. In text feature extraction first required the preprocessing of text data. The preprocessing of text data removal all the unwanted tags colon and syntax used in text data. After that various technique are used for the process of extraction of feature. Some technique is based on transform method, some technique is based on entropy based and some technique is based on frequency of words [9]. After the process of text feature extraction process the categorization and classification process. In this process develop two model of classification process. One model is called training model or learning model and another is called test model. On the basis of training model, the test data are categories. The process of text categorization processes in different scenario model such as complete document and paragraph model and in consequence of sentence. The process of classification depends on the process of user and organization center [11].

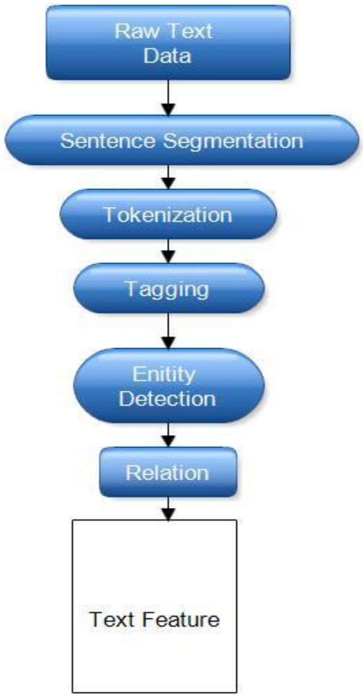

Figure 1: Shows that the process of feature extraction of raw text data.

The extracted text feature pass through ant colony optimization. The ant colony optimization process finds the continuity of words feature. The process of ant colony optimization technique basically describe in terms of artificial ants. The process of ants finds the dissimilar and redundant group of text. The process of feature optimization describe here. The process of feature optimization of multi-level data passes through the feature space of ant colony optimization. The mapping of text data feature attribute according to their artificial ants required some standard derivation and parameter. On the basis of parameter estimate the feature similarity of two different words. And those features are most similar passes through the process of clustering and increase the capacity of text categorization.

Let F is a feature set and N is the total artificial ants and possibility of ant selection is s1,s2…………sn , now find the selection possibility of two ants in given solution is

$$SP(i,j) = \frac{1}{si - sj} \dots\dots\dots\dots\dots\dots.(1)$$

Where si and sj is the dissimilar probability of two different ants. Now estimate the value of appetence of ants is

$$\text{ACP(i+j)}=\frac{\alpha i+\beta i}{N} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots.(2)$$

Where αi and βi is ants whose selection possibility is maximum in terms of another ants the ratio of selection of ants is defined as $\frac{100}{N}$

On the basis of selection possibility estimate the value of artificial phenomenon value

$$\Delta\tau i = \frac{A.si}{ACP(i+j)} \dots\dots\dots\dots\dots\dots\dots\dots..(3)$$

Where A is constant phenomenon value

Now each iteration of pheromone value is increment and decrement according to their selection probability. The derivation of universal appetence probability is

$$p_{ii}^{k} = \begin{cases} [\tau_{ij(t)}] & \alpha.[k_{ij}]\beta \quad if\ j€j \dots\dots\dots..(4) \\ 0 \quad otherwise & \end{cases}$$

Where kij gives the information of heuristic search space and measure the selection possibility of artificial ants And finally getting the optimal word feature of text document for the processing of optimization

VI. PROPOSED ALGORITHM

In this section discuss the proposed algorithm of multi-label text categorization based on clustering and ant colony optimization algorithm. The ant colony algorithm optimized the feature of text data and passes through the mapping of class. In the proposed algorithm there are two section one section is classifier model and other is text model. The classifier model consist of four phase. In first phase data are transforming into ant feature space. In second phase the transferred ant process through cluster mapping phase. The cluster mapping phase generates the predefined class and find the value of matching of features[12,13,14]. The process is same follow for the preparation of test data. The process of algorithm is given below.

Step1. Initially raw text data passes through feature extractor.
Step2. Here show steps of processing of ACO

Initialize each ant's value.

1) Randomly select the feature vector for the process of head and trail ant matrix.
2) Every ant is examined to find the best match feature.
3) The similarity of feature vector is decrease and the number of optimal feature is going to tailor phase.
4) After that the weight value of vector are adjusted and passes through the feature space of cluster map.

The function feature mapping creates forward class for the classification.
input the feature space of class generated feature Matrix.

1. estimate the feature correlation attribute as

$$\mathrm{Rel}(a,b) = \frac{\mathrm{cov}(a,b)}{\sqrt{\mathrm{var}(a)\times\mathrm{var}(b)}}$$ Here a and b the feature vector of feature matrix

2. The estimated correlation coefficient of feature passes through class.

$$x(t) = w0 + \sum_{j=1}^{\text{total data}} wj \exp\left(\frac{-(\text{total} - xj)}{\sigma^2}\right)$$

3. create the relative feature difference value

$$Rc = \sum_{k=1}^{r}\sum_{i=1}^{m}(hi - h)(eik - et)$$

4. After processing of this of feature data creates class.
5. Generate feature mapping of each class according to the unseen data.
6. The classification measures the Similarity and return the equivalent class of data.
7. If the relevant class are not found that the process going again in feature space.

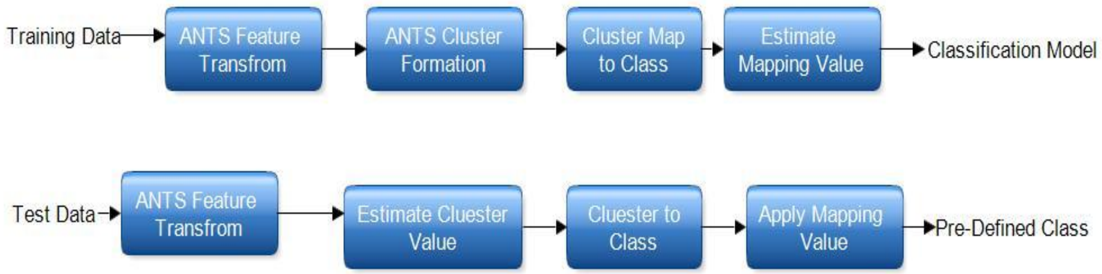

Figure 4.2: Shows that classification model based on ant and clustering technique.

VII. EXPERIMENTAL DETAILS & RESULT ANALYSIS

To evaluate the performance of proposed method of Label Dependency and Feature Similarity Multi Label classification we have use MATLAB software Version 7.8.0 (R2009a)with a variety of Database Dataset used for experimental task.

1. MATLAB

MATLAB is basically a modern programming language environment. The name MATLAB stands for MATrix LABoratory [40]. It has sophisticated data structures, contains built-in editing and debugging tools, and supports object-oriented programming. These factors make MATLAB an excellent tool for teaching and research. MATLAB has many advantages compared to conventional computer languages (e.g. C, FORTRAN) for solving technical problems. MATLAB is an interactive system whose basic data element is an array that does not require dimensioning. The software package has been commercially available since 1984 and is now considered as a standard tool at most universities and industries worldwide. It has powerful built-inroutines that enable a very wide variety of computations. It also has easy to use graphics commands that make the visualization of results immediately available. Specific applications are collected in packages referred to as toolbox. There are toolboxes for signal processing, symbolic computation, control theory, simulation, optimization, and several other fields of applied science and engineering. MATLAB was written originally to provide easy access to matrix software developed by the LINPACK (linear system package) and EISPACK (Eigen system package) projects. MATLAB is a high-performance language for technical computing. It integrates computation, visualization, and programming environment. MATLAB is a tool for numerical computation and visualization. The basic data element is matrix, so if we need a program that manipulates array-based data it is generally fast to write and run in MATLAB. It is a high level language which has many specialized toolboxes for making things easier. MATLAB includes powerful, interactive 2-D and 3-D plotting capabilities that allow one to create revealing color graphics. Advanced visualization tools include surface rendering, lighting, image display and powerful application-specific graphics. The MATLAB family includes support for development of External math- based application. The MATLAB language is designed for interactive or automated computation. The developer of MATLAB , the Mathwork Inc. are no longer supporting Macintosh. Since MATLAB was primarily designed to do calculations- often as fast as C or Fortran. With more than 500 mathematical and engineering functions, MATLAB gives immediate access to high- performance numeric computing. The numerical routines are fast, accurate and reliable. The language includes flow-control, data structures and object oriented programming, plus GUI development tool, debugging features and the ability to link in C, C++ and Fortran routines.

2. DATASET DESCRIPTION

Cleveland

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML research to this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0). Only 14 attribute used.

Table 5.2 Cleveland Dataset values.

Sr No.	Attribute Name
1.	(age)
2.	(sex)
3.	(cp)
4.	(trestbps)
5.	(chol)
6.	(fbs)
7.	(restecg)
8.	(thalach)
9.	(exang)
10.	(oldpeak)
11.	(ca)
13.	(thal)
14.	(num)

Glass data set

This Data Set contains 7 classes, 10 attributes and around 214 instances. Now normalize the dataset, and arrange the data so that new classes appear randomly.

Table 5.3: Shows that original data set of Glass for that data formatted and then used for classification. It is contain 7 class 10 attributes and around 214 instances.

Data Set Characteristics:	Multivariate	Number of Instances:	214	Area:	Physical
Attribute Characteristics:	Real	Number of Attributes:	0	Date Donated	1987-09-01
Associated Tasks:	Classification	Missing Values?	o	Number of Web Hits:	69543

VIII. EXPERIMENTAL RESULT

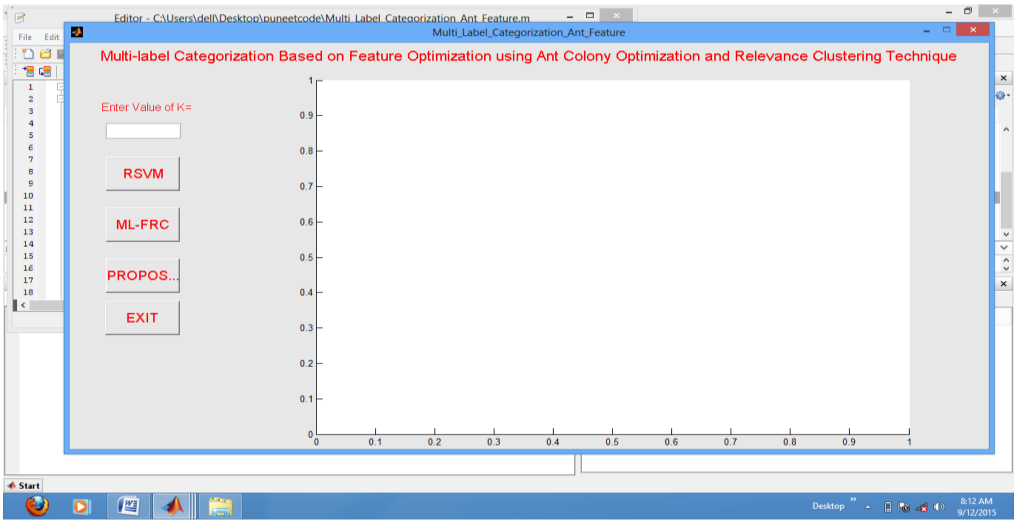

Figure 5.4.1: Show that the Multi-label Categorization Implementation window.

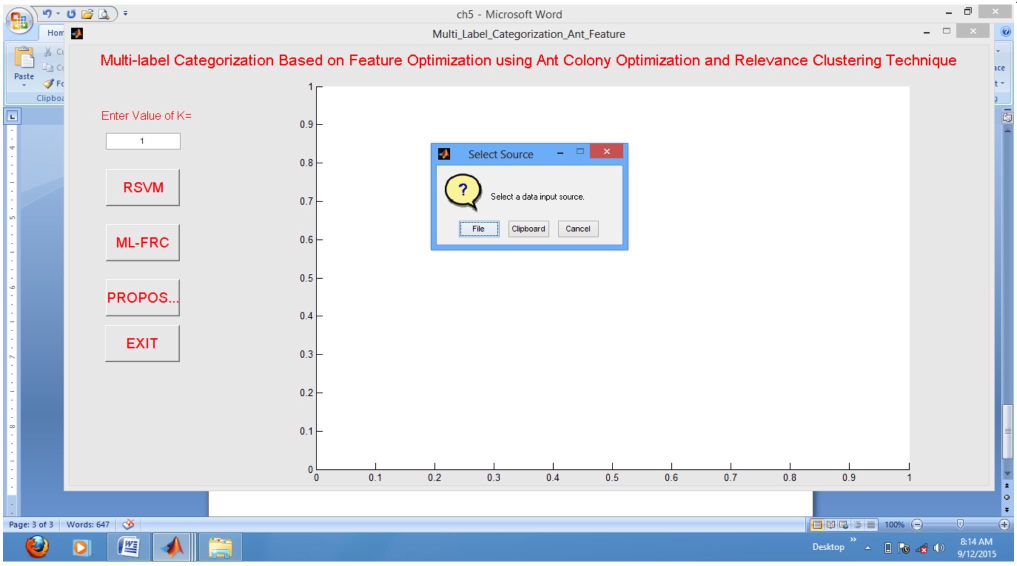

Figure 5.4.2: Show that the Multi-label Categorization window for selection the dataset.

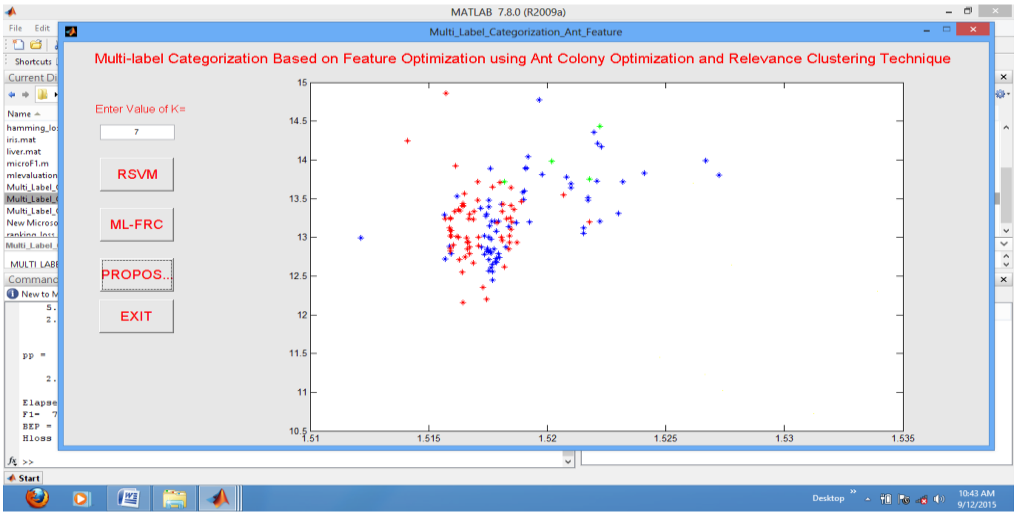

Figure 5.4.25: Show that the Multi-label Categorization for the value of k is 7 using PROPOSED Methods.

IX. COMPARATIVE PERFORMANCE EVALUATION

Table 5.5.7: Show that a Comparative Performance Evaluation for the input value of k is 7 using RSVM, MLFRC and PROPOSED Methods.

Name of Dataset	Input Value of K	Method's Name	F1	BEP	Hamming Loss	Elapsed Time
Glass	7	RSVM	65.0	70.95	16.40	4.80
		ML-FRC	67.0	73.25	13.66	4.86
		PROPOSED	70.0	67.85	10.93	4.69

5.6 Comparative Result Graph

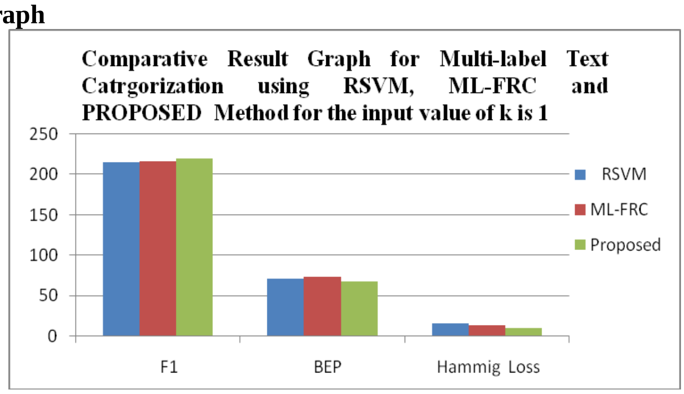

Figure 5.6.1: Show that a Comparative Result Graph for Multi-label Text Categorization using RSVM, ML-FRC and PROPOSED Method for the input value of k is 1, here we show that the our proposed method gives a better result than other existing methods in the terms of result parameters.

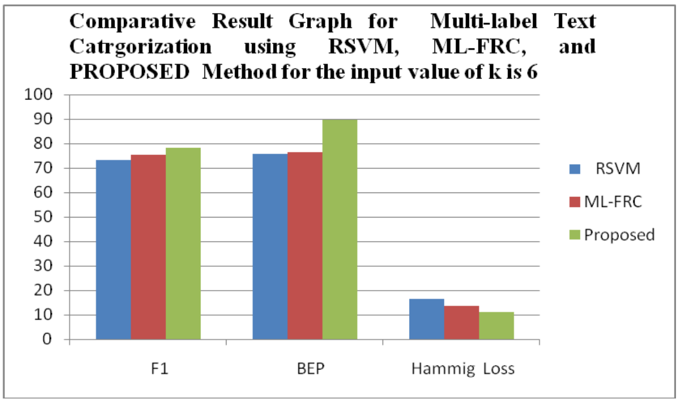

Figure 5.6.6: Show that a Comparative Result Graph for Multi-label Text Categorization using RSVM, ML-FRC and PROPOSED Method for the input value of k is 6, here we show that the our proposed method gives a better result than other existing methods in the terms of result parameters.

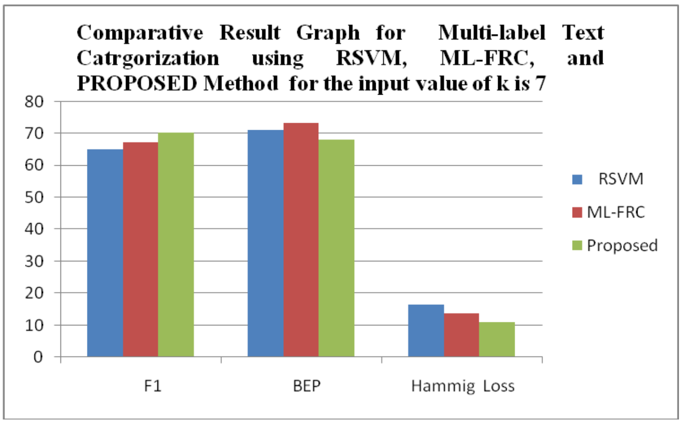

Figure 5.6.7: Show that a Comparative Result Graph for Multi-label Text Categorization using RSVM, ML-FRC and PROPOSED Method for the input value of k is 7, here we show that the proposed method gives a better result than other existing methods in the terms of result parameters.

X. CONCLUSIONS AND FUTURE WORK

Feature optimization and feature selection play an important role multi-label data categorization. In multi-label data categorization multiple features share a common class and the process of classification suffered a problem of selection of relevance feature for the classification. In this paper proposed feature optimization based multi-label data categorization. The process of feature optimization is done by ant colony optimization. The ant colony optimization accrued the relevant common feature of document to class. For the process of classification used cluster mapping classification technique. The feature optimization process reduces the loss of data during the transformation of feature mapping during the classification.

In this dissertation proposed multi-label categorization algorithm based on ANTS algorithm and cluster mapping technique. In proposed algorithm the ants colony optimization algorithm play a task of feature optimization of data mapping of classification. The optimization process also reduces the problem of data dimensions and loss of data. The better selection of feature gives better result of classification and prediction of data categorization.

For the validation of proposed algorithm implemented in MATLAB software and used some standard dataset for the process of classification. All these data obtained from UCI machine learning centre. Our experimental result shows that better classification result instead of RSVM and MLKNN algorithm.

In the direction of future work in multi-label classification technique some bottleneck problem is still remain such as transformation loss of data and dimension reduction. In this direction work is focused on. Also future used in case of real data for the purpose of classification. But the computational time of process is increase. In future we used sampling method for the reduction of time and improvement of minority class classification. And another work of future is RBF sampling process applied in multi-class classification for better mapping of feature space.

REFERENCES

[1] Shie-Jue Lee, Jung-Yi Jiang "Multilabel Text Categorization Based on Fuzzy Relevance Clustering" IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL-22, 2014. Pp 1457-1471.

[2] Francois Queyroi, Maylis Delest, Jean-Marc Fedou, Guy Melancon "Assessing the Quality of Multilevel Graph Clustering" Springer, 2014. Pp 1-20.

[3] Jose Antonio Sanz, Alberto Fernandez, Humberto Bustince, Francisco Herrera "IVTURS: A Linguistic Fuzzy Rule-Based Classification System Based On a New Interval-Valued Fuzzy Reasoning Method With Tuning and Rule Selection" IEEE TRANSACTIONS ON FUZZY SYSTEMS, Vol-21, 2013. Pp 399-412.

[4] Min-Ling Zhang , Zhi-Hua Zhou "A Review on Multi-Label Learning Algorithms" IEEE, 2013, Pp 1-13.
[5] Grigorios Tsoumakas, Affiliate, , Ioannis Katakis, Ioannis Vlahavas "Random k-Labelsets for Multi-Label Classification" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE, 2010. Pp 1-12.
[6] Andrea Esuli , Tiziano Fagni, Fabrizio Sebastiani "Boosting multi-label hierarchical text categorization" Springer, 2008. Pp 1-27.
[7] Xiangnan Kong, Michael K. Ng, Zhi-Hua Zhou "Transductive Multilabel Learning via Label Set Propagation" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Vol-25, 2013. Pp 704-719.
[8] S.Arul Murugan, Dr. P. Suresh "Hybridization Of Em And Svm Clusters For Efficient Text Categorization" International Journal of Innovative Research in Advanced Engineering, 2014. Pp 163-171.
[9] Ying Yu Witold Pedrycz, Duoqian Miao "Multi-label classification by exploiting label correlations" Elsevier ltd. 2014, Pp 2989-3004.
[10] Grigorios Tsoumakas, Ioannis Katakis "Multi-Label Classification: An overview" 2007. Pp 1-13.
[11] Min-Ling Zhang, Zhi-Hua Zhou "ML-KNN:A lazy learning approach to multi-label learning" Elsevier ltd. 2007. Pp 2038-2048.
[12] Krishnakumar Balasubramanian, Guy Lebanon "The Landmark Selection Method for Multiple Output Prediction" International Conference on Machine Learning, 2012. Pp 1-8.
[13] ai, F, Lin, H. T "Multi-label classification with principle label space transformation" Neural Com-putation, 2012.
[14] D. Lewis and M. Ringuette, "A comparison of two learning algorithms for text categorization," in *Proc. 3rd Annu. Symp. Document Anal. Inf. Retrieval*, 1994, pp. 81–93.
[15] T. Joachims, "A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization," in *Proc. Int. Conf. Mach. Learn.*, 1997, pp. 143–151.
[16] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, pp. 81– 106, 1986.
[17] N. Fuhr and C. Buckley, "A probabilistic learning approach for document indexing," *ACM Trans. Inf. Syst.*, vol. 9, no. 3, pp. 223–248, 1991.
[18] J. R. Quinlan, *C4.5: Program for Machine Learning*. San Mateo, CA, USA: Morgan Kaufmann, 1993.

