ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2016.0176

ENHANCED NEIGHBORHOOD NORMALIZED POINTWISE MUTUAL INFORMATION ALGORITHM FOR CONSTRAINT AWARE DATA CLUSTERING

Pushpa C.N[1], Gerard Deepak[2], Mohammed Zakir[3], Thriveni J[4] and Venugopal K.R[5]
[1,2,3,4,5]*Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, India*
E-mail: [1]pushpacn.uvce@gmail.com, [2]gerry.deepu@gmail.com, [3]zakhirocks@gmail.com, [4]drthrivenij@gmail.com

Abstract

Clustering of similar data items is an important technique in mining useful patterns. To enhance the performance of Clustering, training or learning is an important task. A constraint learning semi-supervised methodology is proposed which incorporates SVM and Normalized Pointwise Mutual Information Computation Strategy to increase the relevance as well as the performance efficiency of clustering. The SVM Classifier is of Hard Margin Type to roughly classify the initial set. A recursive re-clustering approach is proposed for achieving higher degree of relevance in the final clustered set by incorporating ENNPI algorithm. An overall enriched F-Measure value of 94.09% is achieved as compared to existing algorithms.

Keywords:

Clustering, Constraint Learning, Normalized Pointwise Mutual Information, Recursive Re-Clustering, SVM

1. INTRODUCTION

Data is extracted and stored at an immense scale; quite often the necessary information or feasible knowledge is "hidden" in the data and is available directly. Data analysts and users' at various levels require several weeks to months for mining and discovering that useful information. Major portion of the data is never subject to analysis and is mostly neglected. To overcome such situations, data mining comes in the upfront. Data mining helps in classifying and segmenting data using Hypothesis Formation. Non-Trivial Aggregation of implicit data that comprises of potentially useful information from a large data repository is termed as Data Mining. Sometimes, data mining can be referred to as exploring and harvesting that unexplored portion of data from a large data repository and transforming the Data into useful information for future analysis. Data mining tasks incorporate several prediction, classification and clustering techniques. These techniques are primarily statistical or based on certain rules and theorems.

Clustering of similar data items is one of the most frequently used data mining strategies. Clustering refers to grouping of data with similar properties which exhibit similar characteristics under a specific label. Clustering is applicable to several heterogeneous domains and is one of the most important and yet a most feasible operation used in Data Mining. Clustering of data has numerous paradigms where it is never visualized as a problem in data mining. Clustering is always viewed as an essential technique for mining useful patterns from large data sets.

In modern times, owing to the diverse growth of data in the Web due to the increased number of users and businesses, data mining plays a significant role. Clustering the increasing volumes of data becomes a tedious task. Automating the clustering mechanism is a huge problem owing to the efficiency and the quick response of the underlying clustering algorithm. Clustering requires learning as an algorithm and the system needs to be trained at different instances to make it more productive and proactive such that it can cluster data items in an organized manner as per the end user specifications.

For strict clustering as per the users' choices, supervised learning is essential for optimal clustering. The problem arises when supervised learning is applied at every instance as it increases the overall complexity of the algorithm. This makes the convergence to optimal solution more difficult. Henceforth, it is never a correct design methodology to use supervised learning for training data items for clustering. The alternative option is using semi-supervised technique for learning but the data items can deviate from the relevant or desired user intentions. This definitely will reduce the overall accuracy of the system. To avoid this, an active constraint learning approach is followed where semi-supervised learning is made intelligent by proposing certain constraints for training. If the constraints are encountered, then the system initiates learning. The constraints are automatically imposed based on the query input by the user.

The constraint checking is generally probabilistic and is referred to as active learning constraints. The constraints in general can either be probabilistic or rule based. Sometimes, hybrid constraints are also considered. A normalized point wise mutual information computation approach is proposed for clustering of data with respect to dynamic deviation computation among the different items of data. To accelerate the learning process initially a hard margin SVM classifier is used to roughly classify the data. The combination of Support Vector Machines with Normalized Point Wise Mutual Information makes the probability of data relevance very high. The constraint set is recursively re-iterated such that the relevance of the data items is increased.

Motivation: The absence of repetitive re-clustering in the clustered constrained data items creates a void and reduces the overall efficiency and the relevance of the relevance of the clustered dataset. Although most of the clustering mechanisms are user driven, they lack the essence of active learning especially that of constraints and do not respond as expected to the learning environment. They cause latencies and delays to semi-supervised learning.

Contribution: A novel strategy that combines SVM with a probabilistic methodology to actively learn constraints in semi supervised learning environment if proposed. The strategy makes the clustered set highly relevant as recursive re-clustering based validation is carried out and is fed back into the training algorithm. A Pointwise mutual information computation methodology is proposed.

Organization*:* The paper organization is as follows. The section 2 provides a brief overview of related research work. Section 3 presents the Proposed System Architecture. Section 4

describes the detailed Implementation. The Results and Performance Analysis is discussed in section 5. Section 6 concludes the paper.

2. RELATED WORKS

Zeng et al. [1] proposed pairwise constraint clustering technique that is similar to a K-means clustering or spectral clustering approach for improved performance. Moreover, the pairwise constraints are mostly incompatible when it is tried to be imposed under the criteria of maximum margin clustering (MMC) technique which enhances the capability of performance in maximum margin. The core concept followed here is cluster pair wise the learning constraints using the concept of a maximization of the margin and Optimal Loss Function is used to reduce the chances of misinterpreting the constraints. As a result of the outcome of task optimization the non-convex tasks are transformed into quadratic convex tasks.

Thiago et al. [2] has proposed algorithms for online constraint based learning, an online vector quantization approach is proposed for the algorithms that incorporate online learning that is dynamic to incorporate constraint based learning via experiments with nine distinct datasets which overcome batch processing of the individual data items. Furthermore, this class of dynamic clustering reduces the optimal convergence of the solution set at a quicker pace. This makes this approach highly tentative and situational to the data set that has been picked up by the algorithm at an instance of time.

Xu et al. [3] proposed a Web usage clustering, which acknowledges a class of users surfing the same kind of patterns reflecting worthy information to custom Web services. Users of the World Wide Web are clustered with respect to K-Means algorithm on the criteria gathered by Web logs. For a group of users and their respective Web usage history data, their actions are captured and clustered. The test results reflect on robustness and optimization of the K-Means algorithmic product for diverse customized Web based applications.

Raja et al. [4] proposed an improvement to fundamental content based image retrieval (CBIR) method with help of indexes by using K-means data clustering. The Improved quality helps in extracting images from immense databases swiftly. A Cluster index is applied on image data sets with respect to clustering algorithm. For processing, cluster phenomenon utilizes attributes such as pixels, density of color, kind of shape, relevance to feedback as well as wavelet referenced histogram technique to capture coincidence amongst images. Based on the criteria of similarity index the individual images at first clustered before they were checked with the database for matching. The images belonging to these groups were separately classified based on image based content checking.

Sudhir [5] has proposed Data and Image characters and patterns which are vaguely gathered in images, is a combination of various fields that merges methods such as visionary computer, digital processing of images, mining of data, database systems and AI. The Significant function of mining is to produce all major traces without prior knowledge of data patterns. Extraction of data has been done with respect to collaborated collections of images and coherent data.

Bhateja et al. [6] suggests Data Mining when limited to multimedia is said to be Web Image extraction or Mining. Surface of social networking Websites has brought about increase in solving immense measure of multimedia data. Image mining technique involves categorization and clustering. Exercising Data Mining includes gaming, businesses, scientific research and engineering. Image Mining employs a critical role in the audio and visual apps. Existing data in the world is predicted to growth extensively thus accelerating the job sector of data analysts in coming days.

Kaur [7] presented a summary of operation of content based image retrieval systems (CIBR) as to how Hierarchical clustering and K-Means technique can support in efficiency as well as quality to utilize immense data sets, assisting faster image retrieval and also search for most relevant images in huge image database . Apart from faster image retrieval it also allows the search for significant and right images in immense image database. K-means is a clustering technique in accordance to the optimization of final calculation of clustering value that is significant for its robustness in generating accurate results in image extraction. By utilizing k-means users will be able to choose the finer class of image with result those are yielded faster.

Devasena et al. [8] proposed an approach for mining images based on a specific matching technique using artificial neural networks. The database utilized in this experimentation includes distinct types of images. A trial image in this group is a part of a specific source image. The proposed system is evaluated with the images in the original image database and the newly emerged ones. On comparing with number of false extracts with correct extracts, image mining system based on the strategy of ANNs provided results which were best-in-class when it was implemented.

Dhonde et al. [9] have proposed a technique involving proper combination and parameterization based on hidden knowledge, data of the image and similar pattern attributes for better image retrieval. Images are growing in the Web which in turn is responsible for making immense image databases, consequently, to obtain specific image it takes quite a bit of time. So the hindrance is not technology it fundamentally an immense database system. To lower such a database for desired images hierarchical algorithm is suggested which provides the optimal result for an ample dataset. Image pre-processing with respect to features supports to lower the dataset which leans towards efficient retrieval as well as ensures the clear image extraction by incorporating k-means algorithm.

Yi et al. [10] presented a non-static semi-supervised clustering algorithm technique which could effectively update categorization results provided based on newly accepted constraints. The main purpose is to make use of the dynamic clustering methodology incorporated to a search problem with respect to a viable clustering space formulated as a convex-hull based on multiple portioning methods. On the basis of sequentially organized pair-wise constraints, an updating scheme is designed to streamline partitions of data in a precise and accurate manner.

Lu et al. [11] proposed a key idea to generate a probabilistic framework, which incorporated classification preferences such as pairwise constraints that can be considered as behavioral

observation. This framework facilitates to build a discriminative-model with respect to semi- supervised clustering algorithms. This Discriminative model provides a manner to code the indefiniteness of Knowledge in accordance with pair-wise constraints on distinct data sets when compared to constrained conventional clustering techniques, which can attain good clustering with effective fewer pair-wise relations.

Morsier et al. [12] regards the necessity of un-labelled data items to operate Semi-Supervised Novelty detection that could be known as an imbalance problem for classification, solved using Cost-Sensitive Support Vector Machine (CSSVM), but a heavy parameter search is needed. Entire solution path algorithms are proposed for utilizing CS-SVM in order to ease and gain choosing parameters for SSND

Nayak et al. [13] presented a survey on SVM role in distinct data mining problems such as categorization of data, data clustering, analysis prediction and additional applications. On a bigger note, a lot of research has been carried out and contributed in diverse journals for mining of data as well as provided a certain no. of flaws of SVM. The goal of this paper is to generalize discrete extents of SVM for acknowledgement and study, while providing enthusiasts a novel representation of the core in theoretic and applied manner.

Xiong et al. [14] have done a detailed study of querying the best relative contrasts to gain real learning with slightest user strength. To provide an essential class technique that is enforced by the user to offer such constraints. An information based principle that chooses a triplet whose result points to the most probable information improvement among the modules of a set of samples. Trials demonstrate the projected technique constantly.

Savitha [15] has proposed a semi-supervised clustering methodology that is a resultant of pairwise constraints. These constraints depend on two factors which include must-link constraint and cannot-link constraint. The proposed iteration based framework incorporates sequential naive batch processing to enhance the clustering outcome. This framework needs recurrent re-grouping of data with an increasingly maturing constraint set. To report this increasingly maturing constraint set, a batch technique is used that extracts a degree of points grounded based on querying in each iterations'.

Tjandrasaet al. [16] have formulated a study to categorize the reasonable and strict ordering of non-incremental diabetes based retinal fundus in imaging. The hard flow is partitioned by utilizing the K-means form of clustering. Partitioned areas are gathered to acquire a characteristic vector that includes these regions, the boundaries, the amount of centroid and respective standard deviation associated with it. Three distinct classifiers are used which are Support Vector Machine that is soft margin based Perceptron, which is multi-layered and Radial Basis Network, an accuracy of 89%, 91%, and 85% is said to be accomplished respectively.

Manimala et al. [17] have proposed a data collection algorithm for recognizing substantial training data exercised for organizing power feature proceedings. The goal of this paper is to minimize execution time, complexity and improve precision of the standing power feature organization system by decreasing the amount of support vectors. The proposed technique recognizes significant training data and discards inappropriate ones by means of a fuzzy C-means data clustering technique based collection algorithm thus decreasing time of arrangement and improving correctness. Substantial characteristics from raw power quality data are mined using distinct wavelet transformation functions and key training data is identified.

Aravinthan et al. [18] have studied recently proposed clustering algorithms. The key aspect of the survey is to provide a significant review of distinct clustering algorithms in data mining. This paper also suggests improvements that can be made in the proposed clustering algorithms thereby providing a base for researchers to read.

Van der Merwe et al. [19] proposed a technique which involves enhancement of K-means algorithm by utilizing particle swarm optimization and the initial set of cluster mid-points. The Experimental calculations reflect that the improved k-means clustering procedure has significant advantage during the time of execution.

3. PROPOSED ARCHITECTURE

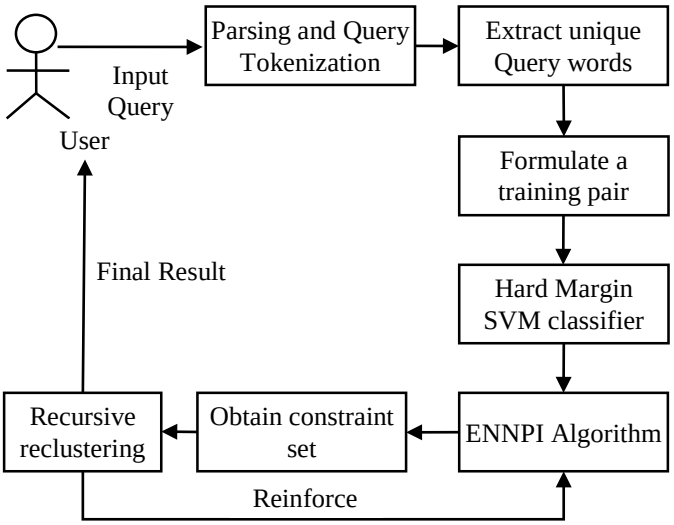

Fig.1. System Architecture

The architecture of the proposed system is represented in Fig.1 where in the User inputs the Query for extracting the useful patterns. The Input Query is initially parsed through a query parser which is further tokenized. At this stage, the input query is decomposed into smaller fragments and individual elements or tokens from the parsed query are extracted. Once the elements are obtained the unique query words are extracted. A random function is chosen for formulating the training pair. The training pair though selected randomly is compared with its relative query specific Pointwise mutual information for relevance. Only if the chosen training pair is highly relevant to the query, then the actual reference training pair is formulated for our active learning.

The constraints are recognized as a hard margin SVM classifier. The hard margin SVM is incorporated owing to the fact that its implementation is feasible. However, the hard margin SVM only checks for boundary values before the data items are actually partitioned through the property of linear seperability. The basic SVM classifier is implemented and its output data items which are unclearly classified are fed into the Proposed Enhanced

Neighborhood Normalized Pointwise Mutual Information (ENNPI) algorithm. ENNPI algorithm is an environment aware framework which categorizes the data items specific to the training pair of elements with a high degree of relevance with respect to data input sets.

The algorithm is a feedback type algorithm where the constraints are examined to classify the data items into an output dynamic constraint set. These constraints contain data items that are classified. The constraint set is further compared with the training pair for its Pointwise mutual information value and is recursively fed into the algorithm for re-clustering. The recursive re-clustering is repeated until there is a convergence to an optimal solution. This is how the active semi-supervised learning based on constraints is obtained. The Pointwise Mutual Information Computation is generally of Normalized form that is calculated using the Eq.(1) and Eq.(2).

$$npmi(x;y) = \frac{pmi(x;y)}{-\log[(x,y)]} \quad (1)$$

$$pmi(x;y) = h(x) + h(y) - h(x,y) \quad (2)$$

The active learning for semi-supervised clustering speeds up the overall clustering process. Apart from attaining a quick response time, continuous supervision of learning the training set is not necessary. Incorporation of a hard margin SVM brings about a directionality for clustering the data items by segregating the data items into vague rough sets which reduces the Pointwise mutual information calculation by an average of fifty percent at every individual stage. The inclusion of re-clustering and recursive framework for processing the partially clustered data items brings about much higher accuracy and higher relevance in clustering of the data items. Moreover, the overall time taken for learning based on the training sets is minimized to a very large extent. The dynamic formulation of training pair and computation of Pointwise mutual information dynamically for a pair of training set of data makes this approach to stand out while reducing the algorithm complexity drastically to a large extent.

4. IMPLEMENTATION

The implementation of the entire system is done using JAVA as the programming language, Netbeans as the IDE on a reliable desktop machine. To incorporate the SVM, no separate tool was used but SVM libraries were incorporated into the backend design. The "SelfOptimizingLinearLibSVM" library was imported to implement the SVM classifier in JAVA platform. Java Swings Framework was used for designing of the front end and the overall design palette. The design pattern which was used is the Factory Design pattern for its ease in usability.

The data sets used were extensive and collected from several sources. Various data sets used were crawled from UCI data repository. The data was structured using MYSQL as a backend data base. The number of records used at every instance was variable during several phases of the experimentation. A few sets of data were also considered from several medical and physics related research data repositories. The proposed algorithm is depicted in Table.1.

Table.1. ENNPI Algorithm

Input: The Set of Data Items D_s, Total Number of Classes K.
Output: Clustered Data Set D_c that is clustered into K Classes.
Step 1: Initialize K=0
Step 2: Select r randomly in set D_s.
Step 3: Input a Query Q.
Step 4: Tokenize Q using StringTokenizer
Step 5:if tokens! =0
Set T_k=tokens
end if
Step 6:for every instance of t of T_k and d of D_s
$pmi(t,d) = h'(t) + h'(d) - h(t, d)$
$npmi(t,d) = pmi(t,d)/-\log[(p(t,d)]$
end for
Step 7:if ($npmi$<0.25)
Cluster d in R_c
Update K
d++
else
repeat Step 5
end
Step 8: The classification of K data sets is obtained.
Step 9: Stop

5. RESULTS AND PERFORMANCE ANALYSIS

The experimentation was carried out on the data sets along with their specific classes as described in Table.1. Standard Formulae were used for Calculation of the *Recall*, *Precision*, *F-Measure* and *Accuracy* of the overall system. The number of Classes, Features and Examples are also indicated in Table.1.

Table.2. Data Sets used for Experimentation

Data Set	Number of Classes	Number of Features	Number of Examples
Wine	4	42	201
Segment	12	22	2736
Glass	12	17	196
E-Coli	7	8	372
Parkinson's	5	9	516
Radio-diagnosis	11	15	986
Pediatrics	4	18	1024
Anatomy	4	14	2879
Laryngology	7	17	2896
Dress	9	16	2146

Cars	8	15	2138
Spectrum	12	22	2896
Cement	5	21	817
Insurance	7	19	442
Sparkles	8	16	1985
Television	12	15	548
Robot	11	12	2887
Lamp	8	11	1547

The *Recall*, *Precision a*nd *F-measure* are tabulated in Table.2. Apparently the performance evaluations are carried out based on these parameters. Since, the clustering is carried out repeatedly until the convergence of optimal solutions the proposed system has a higher precision, recall and f-measure than that of the existing systems. The Precision is defined as in Eq.(3) and Recall is depicted in Eq.(4).

$$\text{Precision}=\frac{\text{No. of data items retrieved and relevant}}{\text{Total No. of data items that are retrieved}} \quad (3)$$

$$\text{Recall}=\frac{\text{No. of data items retrieved and relevant}}{\text{Total No. of data items that are relevant}} \quad (4)$$

The F-Measure F is depicted as the harmonic mean of the Precision and the Recall R and is given by Eq.(5).

$$\text{F-Measure}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \quad (5)$$

Accuracy depicted in Eq.(6) is defined as the average of the sum of recall and precision.

$$\text{Accuracy}=\frac{\text{Precision}+\text{Recall}}{2} \quad (6)$$

From Table.2, one can definitely infer that the average *recall*, *precision* and *accuracy* is 94.19%, 93.99% and 94.09% respectively, which is definitely a high performance index for such systems which employ semi-supervised clustering for learning constraints actively based on an input query. The proposed algorithm performs to its best for data sets like Parkinson's and Lamp but performs at an average rate for a data set like Sparkles and doesn't perform well in the case of a data set like segment. The reason is due to the organization of the features and the individuals of in the related data set. The individuals in Segment data set are actually unrelated. This was analyzed on the analysis of the data sets and semantic heterogeneity between the examples or individuals yields a large value. This is the reason why Segment doesn't perform ideally.

Table.3. Performance Metrics of the Proposed System

Data set	Recall %	Precision %	Accuracy %
Wine	95.67	94.36	95.01
Segment	78.67	77.54	78.10
Mass	96.23	95.74	95.99
E-Coli	97.46	96.36	96.91
Parkinson's	98.36	98.21	98.26

Radio diagnosis	97.54	96.45	96.98
Pediatrics	92.63	91.35	91.98
Anatomy	95.82	94.86	95.36
Laryngology	94.32	96.53	95.42
Dress	93.81	92.62	93.21
Cars	91.52	92.59	92.05
Spectrum	89.91	90.02	89.96
Cement	95.82	95.58	95.70
Insurance	96.33	96.84	96.58
Sparkles	88.68	89.92	89.30
Television	97.34	97.83	97.58
Robot	97.55	97.63	97.59
Lamp	97.89	97.46	97.67

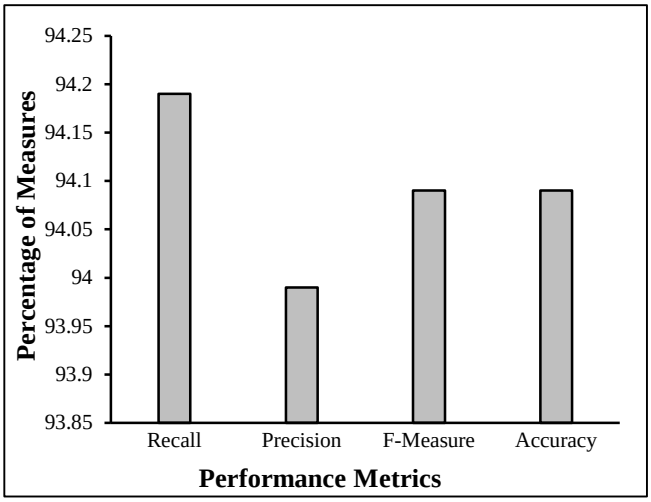

Fig.2. Average percentage of performance measures

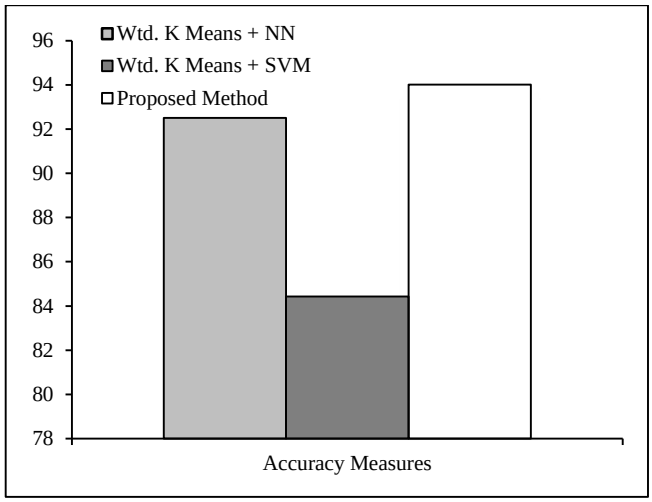

Fig.3. Performance Comparison of the Proposed Method

The average *F-Measure* percent of our system is 94.09 which outperform many existing systems. The Fig.2 depicts the average *recall*, *precision*, *accuracy* and *F-Measure* of the proposed

system. The average accuracy of two existing approaches namely the Weighted K-Means with Neural Networks [20] and Weighted K-Means with SVM [20] is compared with the Proposed Method and is depicted in Fig.3. It is clearly evident that the proposed strategy is much better than the existing methods. The main reason for achieving a higher accuracy value is the inclusion of Normalized Pointwise Mutual Information computation strategy that is primarily a semantic technique which is used for Clustering similar data.

6. CONCLUSION

In this paper, a novel methodology that combines SVM and Point Wise Mutual Information Strategy is implemented for incorporating Semi-Supervised Clustering. The strategy is implemented as ENNPI algorithm which is environment aware and adapts to the learning based on the changing environment factor. The SVM classifier which is incorporated is of hard margin type which initially classifies the data items into vague sets which is further clustered increasing the overall performance of the algorithm. A recursive re-clustering technique is also included to enhance the relevance of the clustered data items. The proposed approach is driven by the input query submitted. A future enhancement can be reducing the number of comparisons of Pointwise mutual information function and further increase the relevance of results. An overall *F-Measure* of 94.09% is achieved.

REFERENCES

[1] Hong Zeng and Yiu Ming Cheung, "Semi-Supervised Maximum Margin Clustering with Pairwise Constraints", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, No. 5, pp. 926-939, 2012.

[2] Thiago F Covoes, Estevam R. Hruschka and Joydeep Ghosh, "Competitive Learning with Pairwise Constraints", *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 24, No. 1, pp. 164-169, 2013.

[3] Jinhuaxu and Hong Liu, "Web User Clustering Analysis based on K-Means Algorithm", *Proceedings of IEEE International Conference on Information Networking and Automation*, Vol. 2, pp. 2-6, 2010.

[4] N.V. Murali Krishna Raja and K. Shirin Bhanu, "Content Bases Image Search and Retrieval Using Indexing By K-Means Clustering Technique", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 2, No. 5, pp. 2181-2189, 2013.

[5] Ramadass Sudhir, "A Survey on Image Mining Techniques: Theory and Applications", *Computer Engineering and Intelligent Systems*, Vol. 2, No. 6, pp. 44-52, 2011.

[6] Preetika Bhateja, Pinki Sehrawat and Avdesh Bhardawaj, "An Analysis of Data Mining, Web Image Mining and their Applications", *International Journal of Information and Computation Technology*, Vol. 3, No. 6, pp. 603-608, 2013.

[7] Gurpreet Kaur, "A Review: Content Base Image Mining Technique for Image Retrieval using Hybrid Clustering", *International Journal of Advanced Research in Computer Engineering and Technology*, Vol. 2, No. 6, pp. 1974-1978, 2013.

[8] C. Lakshmi Devasena and M. Hemalatha, "A Hybrid Image Mining Technique using LIM-based Data Mining Algorithm", *International Journal of Computer Applications*, Vol. 25, No. 2, pp. 11-15, 2011.

[9] Dhonde Parag and C.M. Raut, "Precise and Proficient Image Mining using Hierarchical K-Means Algorithm", *International Journal of Scientific and Research Publications*, Vol. 5, No. 1, pp. 1-4, 2015.

[10] Jinfeng Yi, Lijun Zhang, Tianbao Yang, Wei Liu and Jun Wang, "An Efficient Semi-Supervised Clustering Algorithm with Sequential Constraints", *Proceedings of the 21st ACM International Conference on Knowledge Discovery and Data Mining*, pp. 1405-1414, 2015.

[11] Zhengdong Lu, "Semi-Supervised Clustering with Pairwise Constraints: A Discriminative Approach", *Proceedings of 11th International Conference on Artificial Intelligence and Statistics*, pp. 299-306, 2007.

[12] Frank De Morsier, Devis Tuia, Maurice Borgeaud, Volker Gass and Jean Philippe Thiran, "Semi-Supervised Novelty Detection using SVM Entire Solution Path Geoscience and Remote Sensing", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 51, No. 4, pp. 1939-1950, 2013.

[13] J. Nayak, B. Naik and H.S. Behera, "A Comprehensive Survey on Support Vector Machine in Data Mining Tasks: Applications and Challenges", *International Journal of Database Theory and Application*, Vol. 8, No. 1, pp. 169-186, 2015.

[14] Sicheng Xiong,Yuanli Pei, Romer Rosales and Xiaoli Z. Fern, "Active Learning from Relative Comparisons", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 27, No. 12, pp. 3166-3175, 2015.

[15] S. Savitha and M. Sakthi Meena, "An Efficient Iterative Framework for Semi-Supervised Clustering Based Batch Sequential Active Learning approach", *International Journal On Engineering Technology and Sciences*, Vol. 2, No. 4, pp. 27-31, 2015.

[16] Handayani Tjandrasa, Isye Arieshanti, Radityo Anggoro, "Classification of Non-Proliferative Diabetic Retinopathy based on Segmented Exudates using K-Means Clustering", *International Journal of Image, Graphics and Signal Processing*, Vol. 7, No. 1, pp. 1-8, 2014.

[17] K. Manimala, I.G. David and K. Selvi, "A Novel Data Selection Technique using Fuzzy C-means Clustering to Enhance SVM based Power Quality Classification", *Soft Computing – A Fusion of Foundations, Methodologies and Applications*, Vol. 19, No. 11, pp. 3123-3144, 2014.

[18] K. Aravinthan and M. Vanitha, "Literature Survey on Clustering Algorithms", *International Journal of Advanced Research in Computer Science and Management Studies*, Vol. 3, No. 4, pp. 447-453, 2015.

[19] D.W Van Der Merwe and A.P. Engelbrecht, "Data Clustering using Particle Swarm Optimization", *Evolutionary Computation*, Vol. 1, pp. 215-220, 2003.

[20] Palwinder Kaur, Usvir Kaur and Dheerendra Singh, "Hybrid Clustering and Classification Using Weighted K Mean Neural Networks and SVM", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 4, No. 9, pp. 670-676, 2014.