EFFICIENT DOMAIN DRIVEN DATA MINING BASED ON FUZZY

Pankaj Jain[1], Surendra Singh Chauhan[2]
[2]AP, [1,2]Computer Science Engg, Siddhi Vinayak College of Science & Higher Education, Alwar Rajasthan

ABSTRACT: This paper actualizes the D3m based approach for execution checking of workers of an association on MATLAB stage. It is obviously understood that there is a requirement for space driven information mining, and endeavors are required to create relating methods and applications. The innovative work is required for finding noteworthy learning from complex area issues, upgrading cooperation and lessening the hole amongst the scholarly world and business, and driving an outlook change from intriguing shrouded design mining to significant information revelation in shifting data mining domains. The existing work uses the k-mean clustering with the SVM classifier data driven mining. The paper modifies the existing technique by using BBO optimization along with k-mean clustering and SVM classifier with fuzzy to classify the dataset. The results show the significance of the technique.

Keywords: SVM, k-mean clustering, BBO, Fuzzy, D3M.

I. INTRODUCTION

Data Mining is a strategy for removing data that decides and break down specific information qualities. In D3M universal insight is joined into the mining procedure and models, and a comparing critical thinking framework is shaped as the space for learning revelation and conveyance [1][2]. D3M philosophy will have the capacity to cook for hierarchical variables, client inclinations and business needs. This review gives a brought together space Driven Data Mining (D3M) approach for assessing information insight, area knowledge, human knowledge, arrange insight, social insight, and meta amalgamation of universal knowledge in business associations like IT Industries. This review analyzed supposition mining of virtual colleagues as subjective measure for their execution assessment framework[3]. Knowledge Discovery from Data (KDD) is a standout amongst the most dynamic regions in Information Technology A review of information digging for business applications has demonstrated that there is a major crevice between scholarly destinations and business objectives, and between scholastic yields and business desires[4]. Conventional information mining research for the most part spotlights on creating, illustrating, and pushing the utilization of particular calculations and models. The procedure of information mining stops at example recognizable proof. Subsequently, a broadly observed certainty is that 1) numerous calculations have been planned of which not very many are repeatable and executable in the genuine world.2) regularly many examples are mined yet a noteworthy extent of them are either realistic or of no specific enthusiasm to business, and 3) end clients by and large can't without much of a stretch comprehend and take them over for business utilize[5][6]. It is seen that the discoveries of KDD are not noteworthy, and need delicate power in taking care of true complex issues. Space driven information mining (D3M) has been proposed to handle the above issues, and advance the outlook change from "information focused learning disclosure" to "area driven, noteworthy learning conveyance."[7][8]. Real-world information mining is a mind boggling critical thinking framework. The fundamental target of D3M is to improve the noteworthiness of recognized examples for critical thinking. The expression "noteworthiness" measures the capacity of an example to provoke a client to take solid activities further bolstering his/her good fortune in this present reality[9][10]. It chiefly measures the capacity to propose business basic leadership activities. Table 1 demonstrates an examination of real parts of Data Driven Data Mining and Domain Driven Data mining. The correlation is done by utilizing perspectives, for example, basis, objective, information, prepare, system, foundation, ease of use and so on[11][12][13].

Table 1. Shows a comparison of major aspects of Data Driven Data Mining and Domain Driven Data mining [3]

Aspects	Domain Driven	Data Driven
Rationale	Data and ubiquitous intelligence disclose problem-solving solutions	Data tells a story
Objective	Effective problem solving	Innovative and effective algorithm
Data	Real-life data and surrounding information	Abstract, synthetic and refined data
Process	Multi-step, iterative and interactive on demand	One-off
Mechanism	Human centered or human-mining cooperated	Automated
Infrastructure	Closed-loop problem solving systems in open environment	Closed pattern mining system
Usability	Ad-hoc, dynamic and customizable models and processes	Predefined models and processes
Deliverable	Business friendly decision support systems	Patterns
Deployment	Well-founded artwork in problem solving	Solid validation
Evaluation	Tradeoff between technical significance and business expectation	Technical metrics

II. OPTIMIZATION

Optimization is a commonly encountered mathematical problem in all engineering disciplines. It literally means finding the best possible/desirable solution. Optimization problems are wide ranging and numerous, hence methods for solving these problems ought to be, an active research topic. Optimization algorithms can be either deterministic or stochastic in nature. Former methods to solve optimization problems require enormous computational efforts, which tend to fail as the problem size increases. This is the motivation for employing bio inspired stochastic optimization algorithms as Computationally efficient alternatives to deterministic approach. There are various algorithms which are inspired from bio a list of some of the algorithms is shown in Table 2

Table 2. Optimization Algorithms

Evolutionary based	Swarm intelligence based	Ecology based
1)GeneticAlgorithm	1) ParticleSwarm Optimization	1) PS20
2)Genetic Programming	2) Ant Colony Optimization	2)Invasive Weed ColonyOptimization
3Differential Evolution	3) Artificial Bee ColonyAlgorithm	
4)Biogeography Based Optimization	4)Fish Swarm Algorithm	

Evolutionary algorithms: EA's are most known, established algorithms among all other optimization algorithms. EA's use the methods used by all living organisms to interact with each other. These algorithms used this powerful strategy to find solution to hard problems. EAs are nondeterministic algorithms or cost based algorithms.

Swarm Intelligence based algorithms: are based on collective behaviour of organisms. SI works on the implementation of groups of simple agents that are based on the behaviour of real world insect swarms, as a problem solving tool.

Ecology based: Natural ecosystems provides rich source of mechanisms for designing and solving difficult engineering and computer science problems. It comprises the living organisms along with the a biotic environment with which organisms interact such as air, soil, water etc. There can be numerous and complex types of interactions among the species of ecosystem. Also this can occur as interspecies interaction (between species) or intra species interaction (within species). The nature of these interactions can be cooperative/ competitive.

III. PROPOSED WORK

This vital activity is important to reduce the evil impacts of a contracting workforce. Authoritative records and feelings assume an indispensable part in any association to accomplish the goals. In our current approach we have k mean for bunching of criticisms and afterward we utilize SVM based classifier to order the information. The inconvenience of k mean is that it is moderate, may meet to an answer that is a nearby least of the goal work. Last order in completed with SVM classifier [14]. The impediment of SVM is its high computational cost and results are very reliant on the preparation so it is tedious approach. So we can state that current work have some detriment, for example, the procedure is tedious and results are not that exact thus we can enhance the entire approach by utilizing more successful and proficient calculations. This will permit us to classify information in more effective way than existing method. By doing this we can see the critical change in computational time and more productive aftereffects of clustering. K-mean grouping have a few detriments and we can enhance them utilizing BBO as a part of mix with k-mean so we can state that utilizing the effective technique for the bunching and fluffy govern sets for arrangement we will have the capacity to defeat o issues of precision, decrease the time traverse and function admirably on enormous and worldwide information sets. So the entire framework can be enhanced and we can get more proficient outcomes.

3.1 WORK METHODOLOGY

- The methodology of the work will require the proposed approach to classify the organizational data and opinion data.
- Decision support System formation is carried out using k-mean clustering, biogeography based optimization and fuzzy logic.
- Use MATLAB based platform to implement proposed approach.

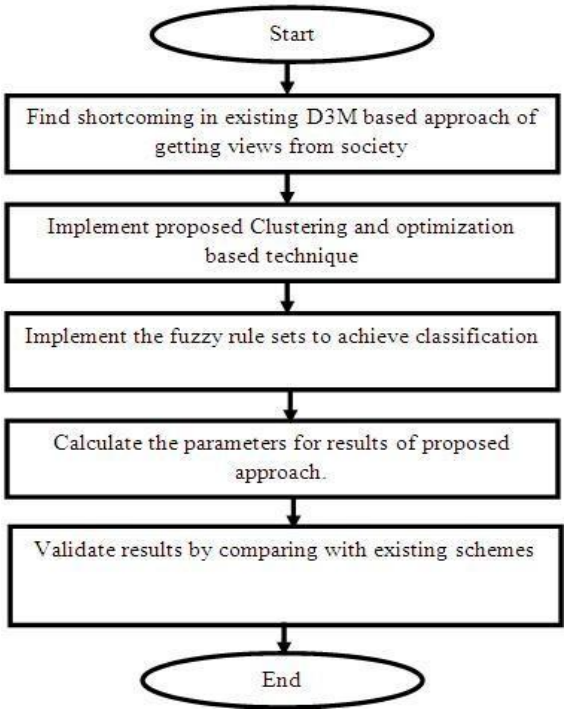

Figure 1 Flow chart of work methodology

Calculated and analyzed these parameters which are described as follows:

- Computation time.
- Specificity.
- Sensitivity.
- Accuracy.

At last, the comparison of existing approach and proposed approach which was implemented in MATLAB platform.

3.2 DATA COLLECTION
Gather inputs, suppositions and remarks as unstructured content from various data sources, for example, criticism online structures, messages, web journals, open discussions which are identified with that business association. Information mining from records and profiles and assessment vault is done in this progression.

3.3 PRE-PROCESSING:
The pre-handling is additionally imperative with a specific end goal to expel pointless words or unimportant words from the client's suppositions. It manages strings tokenization and accentuations evacuation and slangs expulsion. This preparing framework bargains just the portrayal part of every audit, here handling implies part survey into sentences to make a plain content document of surveys. Perform ETL (Extraction, Transformation and Loading) pre-handling to expel clamor from the data sources.

3.4 CLUSTERING:
Clustering the pre-prepared criticism information into important classes by applying K-Means with Biogeography based streamlining. Imagine the classes (i.e. bunches in a high dimensional space to comprehend the

3.4.1 K-MEAN ALGORITHM
K-Means is a parcel based calculation which is one of the easiest unsupervised learning calculations that take care of the notable grouping issue. K mean bunching is most normal kind of centroid based grouping. The technique takes after a straightforward and simple approach to order a given information set through a specific number of groups (accept k bunches) settled from the earlier [10]. The principle thought is to characterize k centroids, one for every group. These centroids ought to be put cleverly in view of various area causes distinctive outcome. Along these lines, the better decision is to place them however much as could reasonably be expected far from each other. The following stride is to take every guide having a place toward a given information set and partner it to the closest centroid. At the point when no point is pending, the initial step is finished and an early gathering age is finished. Now it is important to re-compute k new centroids as banish focuses of the groups coming about because of the past stride. Subsequent to getting these k new centroids, another coupling must be done between similar information set focuses and the closest new centroid. A circle has been created. As an aftereffect of this circle, one may see that the k centroids change their area well ordered until no more changes are finished. At the end of the day centroids don't move any more. The appeal of the k-implies lies in its effortlessness and adaptability. Regardless of different calculations being accessible, k-implies keeps on being an alluring strategy due to its union properties. Be that as it may, it experiences significant deficiencies that have been a reason for it not being actualized on extensive datasets. The most essential among these are:

- K-means is moderate and scales inadequately as for the time it takes for substantial number of focuses.
- The calculation may merge to an answer that is a neighborhood least of the goal work.
- Starting choice of the quantity of bunch must be beforehand known and indicated by the client.
- Comes about straightforwardly rely on upon the underlying centroid of bunch picked by calculation.

The K-implies calculation assembles the arrangement of information focuses in space into a predefined number of bunches. In such manner, the Euclidean separation is ordinarily utilized as a comparability measure. K-means is a grouping calculation that plans to segment the arrangement of perception focuses into K bunches. Give R a chance to be the arrangement of genuine numbers and R^d be d – dimensional vector space. Given R^d is subset of a limited set X={x1, x2 … , xn}, where n is the quantity of vectors. The K-implies calculation segments the set X into subset S, whose subsets are S= {S1,S2,… , SK}, where K is a predefined number. Every group is spoken to by a vector c, C= {c1,c2,… ,cK} is the inside set in the vector space.

3.4.2 BIOGEOGRAPHY BASED OPTIMIZATION
The idea of Biogeography Based Optimization (BBO) was initially exhibited by D. Simon in 2008. It is a populace based transformative calculation that depends on the arithmetic of biogeography. Biogeography is the investigation of topographical dissemination of natural creatures. Biogeography depicts how species move starting with one island/living space then onto the next, how new species emerge, and how species get to be distinctly wiped out. Topographical zones that are most reasonable for organic species gangs high environment reasonableness record (HSI). BBO has certain elements in a similar manner as other science based calculations. Like GAs and PSO, BBO has a method for sharing data between arrangements. GA arrangements "kick the bucket" toward the end of every era, while PSO and BBO arrangements survive always (in spite of the fact that their qualities change as the improvement procedure advances). PSO arrangements will probably bunch together in comparative gatherings, while GA and BBO arrangements don't really have any inherent propensity to group. Above all else the BBO helps in discovering the centroids for bunch formation.BBO helps K-Means in discovering the separation of different items that will be in clusters.BBO helps in consolidating the objects of same component to be in one group. The figure 2 given underneath shows an island relocation display. The migration rate and the resettlement rate are elements of the quantity of species on the island. The greatest conceivable movement rate happens when there are zero species on the island. As the quantity of species builds, the island turns out to be more swarmed, less species can survive migration, and the movement rate diminishes. The biggest conceivable number of species that the living space can support is , and soon thereafter the movement rate is zero. In the event that there are no species on the island, then the resettlement rate is zero. As the quantity of species on the island expands, it turns out to be more swarmed, more species agents can leave

the island, and the migration rate increments. At the point when the island contains the biggest number of conceivable species , the displacement rate achieves its greatest conceivable esteem E.

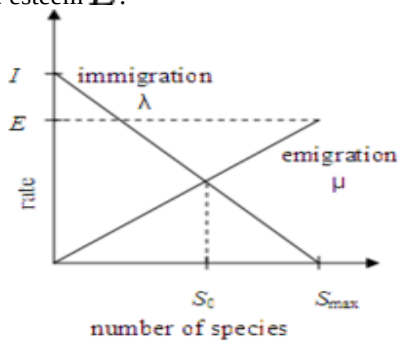

Figure 2 Island immigration model [13]

The BBO algorithm consists of following steps [8]:

1. Initialize the most extreme species check, greatest migration rate, most extreme migration rate, change rate and elitism parameter.
2. Initialize an arbitrary arrangement of environments; every natural surroundings comparing to answer for the given issue.
3. For every environment, outline HSI to the quantity of species, movement rate and displacement rate.
4. Probabilistically utilize migration and displacement to adjust each non-tip top natural surroundings, then recomputed each HSI.
5. For every territory, overhaul the likelihood of its species check. At that point transform each non-first class living space in view of its likelihood and recomputed each HSI.
6. Go to step 3 for the following emphasis. This circle can be ended after a predefined number of eras or after a satisfactory issue arrangement has been found.

BBO additionally plainly varies from ACO, on the grounds that ACO produces another arrangement of arrangements with every cycle. BBO, then again, keeps up its arrangement of arrangements starting with one emphasis then onto the next, depending on relocation to probabilistically adjust those arrangements. BBO likewise has a few elements that are one of a kind among science based streamlining techniques.

3.5 CLASSIFICATION

Opinion words are fluffy in nature. For instance, the words "Decent", "great", and "marvelous" and the limits among them are not clear. Henceforth, Fuzzy rationale can without much of a stretch speak to these sorts of subjective words and appoint to classes with some level of enrollment. This implies these words are as of now in fuzzification arrange. Characterizing fluffy sets for such words should be founded on some master sentiments Since suppositions are fluffy in nature and importance of assessment words can be translated in an unexpected way, Fuzzy rationale is a compelling procedure to be considered here to appropriately remove, break down, arrange and outline conclusions[15].

IV. FLUFFY RATIONALE

Fluffy rationale is a type of numerous esteemed rationale it manages thinking that is surmised as opposed to settled and correct. Fluffy rationale factors may have a truth esteem that extents in degree somewhere around 0 and 1. Fluffy rationale has been reached out to handle the idea of incomplete truth, where reality esteem may go between totally genuine and totally false. Moreover, when etymological factors are utilized, these degrees might be overseen by particular capacities. Nonsensicalness can be depicted as far as what is known as the fuzzjective. Fluffy depends on a hypothesis which relates questions in a set with a level of participation. Fundamental strides which are utilized as a part of fluffy rationale based arrangement are appeared in Figure 3.2.First two stages are as of now performed in the process completed before characterization..

FUZZIFICATION INPUTS: At first the information sources ought to wind up as fluffy information; in our technique we have inputs taking after as: "Decent", "great", and "magnificent" "awful" which are known as assessment words. Unique degree for each of these words are related by human master, for instance: like: 4 love: 5, great: 3, amazing: 6, truly: 5, to a great degree: 9, enjoy:8, exceptionally:

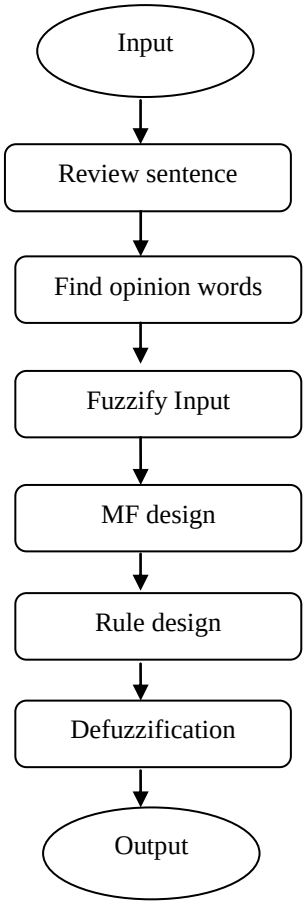

Figure 3 Fuzzy based classification

B. MEMBERSHIP FUNCTION DESIGN: Secondly, membership function (MF) is characterized for discovering participation esteem for each of the data sources. When all is said in done, there are three sorts of MF, to be specific

triangular, trapezoidal, and summed up chime shape. In proposed procedure triangular Membership Function is utilized. Rank of MF is decelerated by human specialists; the phonetic variable used to speak to them was separated into three levels: low, direct and high.

MFs used to display the phonetic marks. Taking after illustrations, when assessment words (i.e. exceptionally, as, appreciate and great) are connected into the triangular participation work (MF), they get these enrollment work values as take after as: μ(very) = 0. 5, μ (like) = 0. 4, μ (extremely) = 0. 9, μ (good) = 0. 3, μ (enjoy) = 0. 8

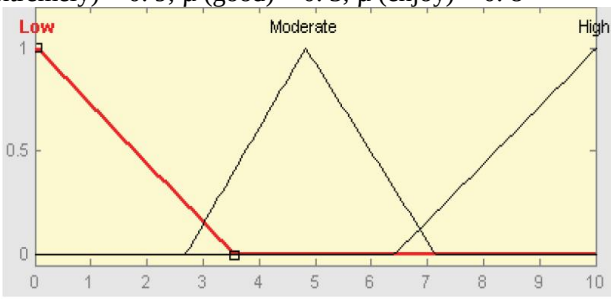

Figure 4 Shows membership function that we use in our method.

C. FUZZY RULES DESIGN: In the third step we characterize a few IF _THEN rules which can bolster the majority of the potential outcomes.

D. DEFUZZIFICATION: To process last yield, defuzzification capacity ought to be utilized to change over them into the fresh esteem and locate the last notion introduction. In this review we utilize Mamdani's defuzzifier, the best known defuzzification administrator in the focal point of gravity defuzzification strategy, which registers the focal point of gravity of the range under the enrollment work:

$$y^* = \frac{\int \mu(y)\, y\, dy}{\int \mu(y)\, dy} \qquad [12]$$

Where y * is the fresh (non-fluffy) esteem; μ(y) is the MF of the relating esteem y in the past outcome. In the third step we define some IF _THEN rules which can support all of the possibilities.

V. RESULT AND DISCUSSIONS

In order to analyze the performance, comparisons will be made with existing approach. Comparisons diagrams will be made based upon the outcomes of the experimental results. Initial parameter will be considered to evaluate the performance of proposed approach. Previous research study of Domain driven data mining approach in data mining is deal with very few applications and parameters for verification and validation of experimental results. But, our research study is based on many important, well known and necessary parameters. An interpretive, phenomenological domain Driven Data Mining (D3M) approach utilizing data mining for Organizational data for objective measurement and opinion mining for subjective measurement enabled a hermeneutic analysis process. The main objective of this research is to investigate the main factors that affect the performance of employees in virtual organization especially IT Companies and to show how these factors can be used for performance evaluation in virtual organization.Performance analysis is done by considering various parametric proofs for each application as a basis of evaluation. In this experiment, dataset is normalized according to the work and used with proposed approach. After applying proposed algorithm on dataset, the performance of system is evaluated and the following are:

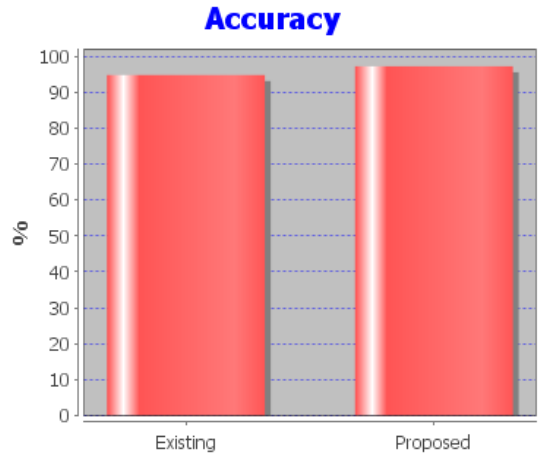

Figure 5 Comparison of accuracy of Proposed and Existing algorithm

In Figure 5 the comparison between existing and proposed work has been shown on the basis of accuracy. It is found that proposed algorithm provide much accurate results as compare to existing approach. The proposed algorithm is more efficient than exiting algorithm.

Figure 6 Computational time taken by the existing and proposed algorithm

As shown in the Figure 6 the comparison between existing and proposed work has been shown on the basis of a Computational Time. It is clear from the figure that by using the proposed approach computational time of the system decreased to a great extent. In this figure the sensitivity of the existing technique is shown. Sensitivity (also called the true positive rate it is the measures of the proportion of actual positives which are correctly identified. In existing technique we have value of sensitivity = 2.828. While In present technique we have value of sensitivity = 3.482 which is greater as compared to existing approach .So in proposed approach true positive rate is high. Specificity (sometimes called the true negative rate) .It is the measures of the proportion of actual negatives which are correctly identified.

In existing technique we have value of Specificity = 4.287. In existing technique we have value of specificity = 5.278. Which is greater as compared to existing approach .So in proposed approach true positive rate is high. The value of both parameters sensitivity and specificity of existing approach is less than as compare to the proposed approach so we can say that our proposed approach is more efficient and effective than the existing approach.

VI. CONCLUSION

The paper concludes that fuzzy based D3M is much efficient than that of SVM based classifier. Result and runtime depends upon initial partition for both of these methods. The advantage of Fuzzy is its low computation cost and it provides more accurate results. This will allow users to categorize data in more efficient manner than existing technique. The Results has been analyzed with the help of parameters and the comparisons are also drawn among the proposed and existing techniques based upon the Computational time taken, accuracy, specificity, sensitivity. In future it can expand the skyline of prominence considerably.

REFERENCES

[1] Al-maimani M., Salim N., Al-naamany A. (2014)," semantic and fuzzy aspects of opinion mining" journal of theoretical and applied information technology,. Vol. 63, No.2 pp. 330

[2] Adali S., Murad M., Kadir R." Sentiment Classification of Customer Reviews Based on Fuzzy logic"IEEE 2010 pp. 1037-1040.

[3] Cao L. (2007), "Domain-driven actionable knowledge discovery," IEEE Intelligent Systems, Vol. 22, No. 4, pp. 78-89.

[4] Cao L. (2010), "Domain-Driven Data Mining: challenges and prospects", IEEE Transaction on Knowledge and Data, Vol. 22, No. 6, pp.765-769.

[5] Cao L., Zhao Y., Zhang C, Yu. P.S. (2010), "Domain driven data mining: D3m methodology ", pp. 27-47.

[6] Cao L. (2007), "Domain-Driven Actionable Knowledge Discovery", IEEE Intelligent Systems, Vol. 22, No. 4, pp. 78-89.

[7] Cao L., Zhao Y., Zhang H., Luo D., Zhang C. (2010), "Flexible Frameworks for Actionable Knowledge Discovery", IEEE Trans. Data and Knowledge Eng., Vol. 22, No.9, pp. 1299-1312.

[8] Cao L. (2008), "Domain Driven Data Mining (D3M)", IEEE International Conference on Data Mining Workshops, pp. 74-76.

[9] Dalal M., Zaveri M. (2014), "Opinion mining from online user reviews using fuzzy linguistic hedges", Hindawi Publishing Corporation Applied Computational Intelligence and Soft Computing, Vol. 2014, No.735942.

[10] Dzogang F., Lesot M. J., Rifqi M., Meunier B. B. "Expressions of graduality for sentiments analysis - A survey"

[11] Elangovan V. R., Ramaraj E. (2013), "Domain driven data mining: An efficient solution for IT management Services on issues in ticket processing", International Journal of Computational Engineering Research, Vol. 03, Issue 5, pp. 31-37.

[12] Ghalib M. R., Vohra S., , Juneja A., "Mining on car database employing learning and clustering algorithms", International Journal of Engineering and Technology (IJET), Vol 5, No 3, pp. 2628-2635

[13] Simon D. (2008), "Biogeography - Based Optimization" IEEE Transactions on Evolutionary Computation, Vol. 12, No.6, pp. 702-713.

[14] Suriyakumari V., Kathiravan A.V (2013), "An ubiquitous domain driven data mining approach for performance monitoring in virtual organizations using 360 Degree data mining & opinion mining", IEEE Pattern Recognition, Informatics and Mobile Engineering (PRIME), pp. 307 – 311.

[15] Tumsare P., Sambare A. S., Jain S. R. (2014), "Opinion mining in natural language processing using sentiwordnet and fuzzy" International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) Volume 3, Issue 3, pp. 154-158.

