Fuzzy Based Data Driven Data Mining Implementation

Kamala kanta Padhi
Raajdhani Engineering College, Bhubaneswar
kamalakanta@rec.ac.in

Abstract

The D3m-based approach to employee performance monitoring within a company is implemented in this study. It is evident that domain-driven data mining is necessary, and efforts must be made to create related methods and applications. Research and development are necessary to drive a paradigm shift from fascinating hidden pattern mining to actionable knowledge discovery in various data mining domains, improve interaction and close the gap between academia and business, and extract useful knowledge from complex domain problems. It is clear from the research above that fuzzy-based D3M outperforms SVM-based classifiers in terms of efficiency.

Keywords: *D3M, SVM, Fuzzy, K-Mean.*

1. Introduction

Data Mining is a method of extracting information that helps to determine and analyze certain data traits [1]. In D3M ubiquitous intelligence is incorporated into the mining process and models, and a corresponding problem-solving system is formed as the space for knowledge discovery and delivery. D3M methodology will be able to cater for organizational factors, user preferences and business needs. This study provides a unified domain Driven Data Mining (D3M) [2] approach for evaluating data intelligence, domain intelligence, human intelligence, network intelligence, social intelligence, and meta synthesis of ubiquitous intelligence in business organizations like IT Industries. This study examined opinion mining of virtual team members as subjective measure for their performance evaluation system.

Knowledge Discovery from Data (KDD) [3] is one of the most active areas in Information Technology A survey of data mining for business applications has shown that there is a big gap between academic objectives and business goals, and between academic outputs and business expectations. Traditional data mining research mainly focuses on developing, demonstrating, and pushing the use of specific algorithms and models. The process of data mining stops at pattern identification. Consequently, a widely seen fact is that 1) many algorithms have been designed of which very few are repeatable and executable in the real world.2) often many patterns are mined but a major proportion of them are either commonsense or of no particular interest to business, and 3) end users generally cannot easily understand and take them over for business use. It is seen that the findings of KDD are not actionable, and lack soft power in solving real-world complex problems. Domain-driven data mining (D3M) has been proposed to tackle the above issues, and promote the paradigm shift from “data-centered knowledge discovery” to “domain-driven, actionable knowledge delivery.” This strategic initiative is necessary to diminish the ill effects of a shrinking workforce. Organizational records and opinions play a vital role in any organization to achieve the goals.In our existing approach we have k mean for clustering of feedbacks and then we use SVM based classifier to classify the data[5]. The disadvantage of k mean is that it is slow, might converge to a solution that is a local minimum of the objective function. Final classification in carried out with SVM classifier. The disadvantage of SVM is its high computational cost and results are highly dependent on the training so it is time consuming approach. So we can say that existing work have some disadvantage such as the process is time consuming and results are not that accurate and so we can improve the whole approach by using more effective and efficient algorithms. This will allow us to categorize data in more efficient manner than existing technique. By doing this we can see the significant improvement in computational time and more efficient results of clustering.

K-mean clustering have some disadvantages and we can improve them using BBo in combination with k-mean so we can say that using the efficient method for the clustering and fuzzy rule sets for classification we will be able to overcome problems of accuracy, reduce the time span and work well on huge and global data sets [7]. So the whole system can be improved and we can get more efficient results.

2. K-Mean Clustering with Fuzzy Classification

2.1 Data Collection

Collect feedbacks, opinions and comments as unstructured text from different information sources such as feedback online forms, emails, blogs, public forums which are related to that business organization. Data mining from records and profiles and opinion repository is carried out in this step.

2.2 Pre-Processing

The pre-processing is also important in order to remove unnecessary words or irrelevant words from the user's opinions. It deals with strings tokenization and punctuations removal and slangs removal. This processing system deals only the description part of each review, here processing means splitting review into sentences to create a plain text file of reviews. Perform ETL (Extraction, Transformation and Loading) pre-processing to remove noise from the information sources.

2.3 Clustering

Cluster the pre-processed feedback data into meaningful categories by applying K-Means with Biogeography based optimization. Visualize the categories (i.e. clusters in a high dimensional space to understand.

2.3.1 K-Mean Algorithm

K-Means is a partition based algorithm which is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. K mean clustering is most common type of centroid based clustering**.**

The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori [6]. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point it is necessary to re-calculate k new centroids as bar centers of the clusters resulting from the previous step. After obtaining these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop, one may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

The attractiveness of the k-means lies in its simplicity and flexibility. In spite of other algorithms being available, k-means continues to be an attractive method because of its convergence properties. However, it suffers from major shortcomings that have been a cause for it not being implemented on large datasets. The most important among these are:

I. K-means is slow and scales poorly with respect to the time it takes for large number of points.
II. The algorithm might converge to a solution that is a local minimum of the objective function.
III. Initial selection of the number of cluster must be previously known and specified by the user.
IV. Results directly depend on the initial centroid of cluster chosen by algorithm.

The K-means algorithm groups the set of data points in space into a predefined number of clusters. In this regard, the Euclidean distance is commonly used as a similarity measure. K-means is a clustering algorithm that aims to partition the set of observation points into K clusters. Let R be the set of real numbers and R^d be d –dimensional vector space. Given R^d is subset of a finite set X={x1, x2 … , xn}, where n is the number of vectors. The K-means algorithm partitions the set X into subset S, whose subsets are S= {S1,S2,…, SK}, where K is a predefined number. Each cluster is represented by a vector c, C= {c1,c2,…,cK} is the center set in the vector space.

The Levenshtein distance is a string metric for measuring the difference between two sequences. The Levenshtein distance between two words is the minimum number of single-character edits (i.e. insertions, deletions or substitutions) required to change one word into the other. It is named after Vladimir Levenshtein, who considered this distance in 1965. Levenshtein distance may also be referred to as edit distance. This is used to calculate the distance in strings in k –mean. It is simplest approach to calculate distance between strings.

Algorithm Steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent the initial group is reached.
2. Calculate the distance between the cluster centre and the data vectors
3. Assign each object to the group that has the minimum distance.
4. When all the objects have been assigned recalculate the cluster center of k centroids.
5. Repeat the steps until the termination condition reached.

Applications of K mean

K-means clustering in particular when using heuristics such as Lloyd's algorithm is rather easy to implement and apply even on large data sets. As such, it has been successfully used in various topics such as ranging from:

1. Market segmentation
2. Computer vision
3. Geostatistics
4. Astronomy
5. Agriculture

K mean is often used as a pre-processing step for other algorithms, for example to find a starting configuration.

2.4 Classification

Opinion words are fuzzy in nature. For example, the words “Nice”, “good”, and “awesome” and the boundaries among them are not clear. Hence, Fuzzy logic can easily represent these types of subjective words and assign to classes with some degree of membership. This means that these words are already in fuzzification stage. Defining fuzzy sets for such words needs to be based on some expert opinions Since opinions are fuzzy in nature and meaning of opinion words can be interpreted differently, Fuzzy logic is an effective technique to be considered here to properly extract, analyze, categorize and summarize opinions.

Fuzzy logic

Fuzzy logic is a form of many-valued logic it deals with reasoning that is approximate rather than fixed and exact. Fuzzy logic variables may have a truth value that ranges in degree between 0 and 1. Fuzzy logic has been extended to handle the concept of partial truth, where the truth value may range between completely true and completely false. Furthermore, when linguistic variables are used, these degrees may be managed by specific functions. Irrationality can be described in terms of what is known as the fuzzjective. Fuzzy is based on a theory which relates objects in a set with a degree of membership. Main steps which are used in fuzzy logic based classification are shown in Figure 3.2.First two steps are already performed in the process carried out prior to classification.

A. Fuzzification Inputs:

At first the inputs should become as fuzzy data; in our method we have inputs following as:

“Nice”, “good”, and “awesome” “bad” which are known as opinion words. Special degree for each of these words are associated by human expert, for example: like: 4 love: 5, good: 3, excellent: 6, really: 5, extremely: 9, enjoy:8, very:

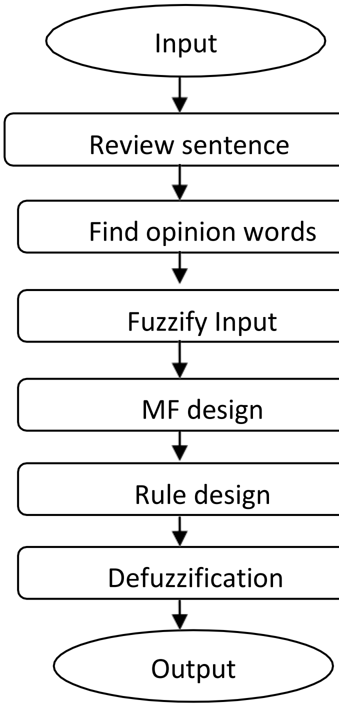

Figure 1: Fuzzy based Classification

B. Membership Function Design:

Secondly, membership function (MF) is defined for finding membership value for each of the inputs. In general, there are three types of MF, namely triangular, trapezoidal, and generalized bell-shape. In proposed technique triangular Membership Function is used. Rank of MF is decelerated by human experts; the linguistic variable used to represent them was divided into three levels: low, moderate and high.

MFs used to present the linguistic labels. Following examples, when opinion words (i.e. very, like, enjoy and good) are applied into the triangular membership function (MF), they obtain these membership function values as follow as: μ(very) = 0. 5, μ (like) = 0. 4, μ (extremely) = 0. 9, μ (good) = 0. 3, μ (enjoy) = 0. 8

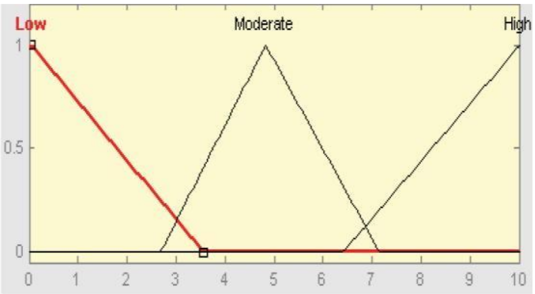

Figure 2: Shows Membership Function that we use in our Method

C. Fuzzy Rules Design:

In the third step we define some IF _THEN rules which can support all of the possibilities.

D. Defuzzification:

To compute final output, defuzzification function should be used to convert them into the crisp value and find the final sentiment orientation. In this study we use Mamdani's defuzzifier, the best known defuzzification operator in the center of gravity defuzzification method, which computes the center of gravity of the area under the membership function:

$$y * = \frac{\int \mu\,(y)y\,dy}{\int \mu\,(y)dy}$$

Where y * is the crisp (non-fuzzy) value; μ(y) is the MF of the corresponding value y in the previous result.

3. Results and Discussion

Experimental results deals with the output. Here, After running the project the next step is to import organizational data which contain ranks given to employees by organization according to their performance and other information such as emp_id, designation, grade, deptt., and experience and opinion data which contain different opinions which are given by different customer on the blog to different employees into the system. These are the files in which the comments are provided by society and the grade and ranking is done by organization. When we import data into system then the system with its pictorial view can be shown. The next step is to pre-process the data which contain comments pre-processing is carried out to remove noise from the data the pre-processing of the file in which the comments are provided by the society is done. In this step the slangs are removed from the comments and any repetition is also removed from the file and all comments are arranged in a proper format so that they are completely understandable. The comments are also arranged in sentence case form so that they became representable. In this step clustering of data is performed with the help of K-mean and BBO. In this step opinion data is categorize in different group depending upon opinion words. We have to categorize data in 5 categories so opinion data is grouped in 5 groups. In this step we have all the comments of a customer together and in the column cluster we have the cluster no at which that comment fall. SVM classify the employees into 5 classes according to their performance. Fuzzy based classifier shows the grade of every employee, employee's unique id designation and IP and then category column which contain classification of employees according to their comments and then combined column shows the final results which is combination of the ranking based and comments based performance. We classify employees in 5 classes' good performer, best performer, average performer, below average and non performer. The results shown by proposed approach are more accurate as compare to the existing approach.

In this work, 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:

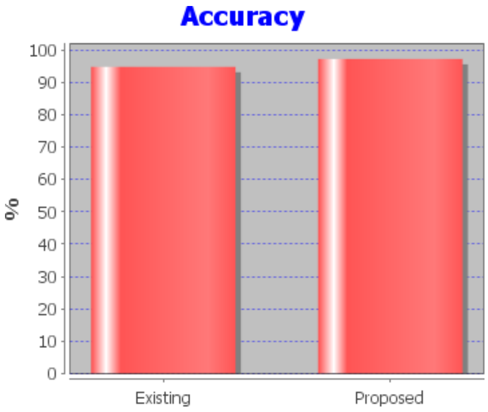

Figure 3: Comparison of Accuracy of Proposed and Existing Algorithm

In Figure 3 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.

Table 1: Result Analysis

Parameter Name	Existing Technique	Proposed Technique
Accuracy	92%	95%
Time Consumed	4520ms	10ms
TP Rate	2.82	3.48

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 TP rate is = 2.828. 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.

4. Conclusion

The 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, TP rate.

References

[1] 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.

[2] 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.

[3] 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

[4] 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.

[5] Saleh M. R., Martin-Valdivia M.T., Montejo-Raez A., Urena-Lopez L.A. (2011), "Experiments with SVM to classify opinions in different domains" Expert Systems with Applications 38, pp.14799–14804.

[6] Sakthi M., Thanamani A. S. (2013), "An enhanced K means clustering using improved hopfield artificial neural network and genetic algorithm", International Journal of Recent Technology and Engineering (IJRTE), Vol. 2, No.3, pp. 16-21.

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