IJCTA, 9(36), 2016, pp. 419-425
© International Science Press

Sentiment Classification using Decision Tree Based Feature Selection

A. Suresh* and C.R. Bharathi**

Abstract : The latest development in access to technology and the Internet, together with the advancement of the Web 2.0 (Social Web), has prompted the introduction of new and intriguing social marvels. From one perspective, the likelihood to express opinion "by anyone, anywhere, on anything", in blogs, forums, review sites has made it workable for people all around the globe to take better and more informed decisions at the time of purchasing products and contracting services. Then again, the companies and public persons are more informed on the impact they have on people, on the grounds that the expansive amount of opinions expressed on them offers a direct and unbiased, global feedback. Sentiment Analysis likewise called "opinion mining" is a sundown sub control of information retrieval and computational historical underpinnings, which is concerned not with the topic a substance, is about, yet rather with the review it passes on. Feature selection has snatched hugeness in view of its diligence to extra classification cost with respect to time and requital load. In this paper, the important center is on feature selection for sentiment analysis utilizing decision trees. The proposed method is evaluated utilizing RatingSystem.com data set, and the exploratory results demonstrate that the proposed feature selection framework is promising.

Keywords: Sentiment Analysis, Inverse Document Frequency (IDF),Leaningr Vector Quantization(LVQ).

1. INTRODUCTION

Humans are social creatures. They can't achieve the level of what we call "human" unless they create in organized social orders, where they are taught standards, rules and laws administering the presence and concurrence of people. Albeit the majority of the times unknowingly, we persistently shape our conduct and attitudes on the premise of these social traditions, of public and private opinions and occasions of the world encompassing us. We give and acknowledge counsel as part of our consistently lives, as part of a custom to knowing, better understanding and coordinating into our encompassing reality. Our responses are based on what we anticipate. Besides, standards of what is permitted and what is not, what is by and large anticipated that would be done or not in a context, trigger our own particular emotional response and our attitude towards the circumstances [1].

Together with the advancement of technology and the developing access to information, we have seen the introduction of another sort of society - that of the connection and communication. In this new context, the part of emotion has become pivotal. Facts determine emotion in people, who assimilate facts and express the impact these facts have on them. Different persons have admittance to these, which they, in their turn, transform affected by their own particular emotional perception. For all intents and purposes, access to information has additionally offered approach to access to emotional response to information, in the light of which information changes. Along these lines, people respond to both facts and attitude on facts. Keeping in mind society is changing, standards are changing alongside it, and world attitude shapes

* Professor & Head, Department of Computer Science and Engineering, Nehru Institute of Engineering and Technology, Coimbatore. Tamil Nadu, India. 641 105

** Associate Professor, Dept. of ECE, Vel Tech University

new standards, under which social orders further change. We can think about various illustrations of facts, on which the general public's and the people's opinion has changed after some time and we can likewise consider facts that stay forbidden subjects in numerous communities around the globe. It is in this manner both fascinating, and in addition testing to see what the opinion is on sure subjects, with the goal that patterns can be anticipated and the right measures taken.

The programmed processing of texts to detect opinion expressed in that, as a unitary collection of research, has been named opinion mining or sentiment analysis. Most work on sentiment analysis has been completed on profoundly subjective text sorts, for example, blogs and product or movie reviews. Creators of such text sorts generally express their opinions uninhibitedly. News articles have gotten a great deal less consideration, in spite of the fact that news predisposition across various news sources has been examined by a couple and some underlying endeavors have focused on sentiment analysis in the news territory [2]. News articles and other media reports normally contain a great deal less plainly expressed opinions. Despite the fact that backing or feedbacks are sometimes expressed, the inclination or sentiment of the journalist is regularly expressed indirectly, for example by highlighting a few facts while perhaps discarding others or by the decision of words.

There are numerous difficulties to Sentiment analysis. The first is an opinion word considered positive in one circumstance and negative in another. The second test is that people express opinions in different ways. Ordinary text processing is on the grounds that constrained contrasts can be recognized between two text pieces, which does not change meaning much. Some research fields are overwhelming in Sentiment analysis: sentiment classification, feature based Sentiment classification and opinion summarization. Sentiment classification classifies whole documents as indicated by opinions to particular objects. In any case, feature-based Sentiment classification considers certain subjects features opinions. Opinion summarization is not quite the same as customary text summarization as the main product features are mined on which customers expressed opinions. Opinion summarization neglects to outline reviews by picking a subset or revises some unique sentences from reviews to capture principle focuses as in customary text summarization.

It is hard for a human reader to find relevant sources, extract related sentences and opinions, read, condense, and organize them into usable forms. Along these lines, automated opinion revelation or summarization frameworks are required. Sentiment analysis likewise called Opinion Mining, originated from this need and is a testing natural language processing/text mining issue. It's colossal value for applications prompted its hazardous development in research, academia and industry [3].

Often utilized data mining dimensionality reduction method is a feature selection that chooses a unique features subset based on particular criteria. It diminishes features number, evacuates irrelevant/redundant/noisy data, giving applications impacts which incorporate accelerating data mining algorithms, enhancing mining performance like predictive accuracy and result comprehensibility. Feature selection is a dynamic research field and created machine learning, and data mining for quite a long time and is presently connected to fields like text mining, genomic analysis, intrusion detection and image retrieval. At the point when new applications developed, numerous difficulties likewise emerged requiring new speculations/methods to address high-dimensional/complex data. Optimal redundancy evacuation, stable feature selection, and auxiliary data and earlier knowledge abuse in feature selection are among the essential and testing issues in feature selection. Progressive, vast volumes of literature were distributed on the research direction of feature selection.

Inverse document frequency (IDF) is an imperative and generally utilized concept as a part of information retrieval. At the point when IDF combines with term frequency (TF), it results in a robust/exceptionally effective term weighting scheme connected across different application territories like databases natural language processing, knowledge management, text classification and information retrieval. There were few endeavors to enhance predetermined number of "classical" IDF formulations chiefly because of the fact that it is nontrivial to change standard IDF formulation in a hypothetically meaningful manner while enhancing effectiveness. There might be heuristic approaches to adjust IDF formulation, however doing as such prompts small understanding with respect to why things moved forward.

Online customer reviews are a critical informative resource valuable for both potential customers and product manufacturers. Reviews are composed in natural language and are sans unstructured texts scheme in website pages. The undertaking of physically scanning gigantic amounts of reviews is computationally difficult and not for all intents and purposes actualized with respect to businesses/customer viewpoints. Subsequently, it is proficient to naturally process different reviews giving important information in a right method. Opinion summarization delivers how to determine sentiment, attitude/opinion a creator expressed in natural language text in regards to a particular feature. A way to deal with mine the product feature and opinion based on both syntactic and semantic information contemplations was proposed in [4]. Use of reliance relations and ontological knowledge with probabilistic based model, demonstrated that this method was more adaptable than others.

[5] Proposed a procedure based on association rule mining to extract product features. The primary thought is that people regularly utilize the same words when they comment on the same product features. At that point incessant itemsets of nouns in reviews are prone to be product features while the occasional ones are less inclined to be product features. This work likewise presented the thought of utilizing opinion words to discover extra (frequently occasional) features. Their algorithm requires that the product class is known. The algorithm determines whether a noun/noun phrase is a feature by computing the pointwise mutual information (PMI) score between the phrase and class particular discriminators, e.g., “of xx”, “xx has”, “xx comes with”, and so on., where xx is a product class. This work initially utilized part-whole patterns for feature mining, however it discovers part-whole based features via searching the Web. Querying the Web is time expending. In our method, we utilize predefined part-whole relation patterns to extract features in a domain corpus. These patterns are domain-independent and genuinely exact.

Taking after the underlying work in [6], a few researchers have further investigated the thought of utilizing opinion words as a part of product feature mining. The extraction rules are composed based on various relations between opinion words and features, and among opinion words and features themselves. Reliance grammar was received to portray these relations. [7] a pattern mining method was utilized. The patterns are relations in the middle of feature and opinion pairs (they call angle assessment pairs). The patterns are mined from a substantial corpus utilizing pattern mining. Statistics from the corpus are utilized to determine the confidence scores of the extraction. All in all information extraction, there are two methodologies: rule-based and factual. Early extraction frameworks are for the most part based on rules e.g., Riloff, 1993). In factual methods, the most well known models are Hidden Markov Models (HMM) (Rabiner, 1989), Maximum Entropy Models (ME) (Chieu et al., 2002) and Conditional Random Fields (CRF) (Lafferty et al., 2001). CRF has been appeared to be the best method. In any case, a restriction of CRF is that it just captures local patterns as opposed to long range patterns.

Previously, researchers grew substantial feature selection algorithms intended for different purposes and every model had its own preferences/detriments. Despite the fact that there were endeavors to survey existing feature selection algorithms, a repository gathering agent feature selection algorithms to encourage comparison/joint study is yet to emerge. To offset this, [8] introduced a feature selection repository intended to gather well known algorithms created in feature selection research to be a platform to encourage application/comparison/joint study. The repository helps researchers accomplish dependable assessment when building up the new feature selection algorithms.

Quicker and accessible web guarantees that people search/learn from divided knowledge. For the most part, colossal volumes of documents and homepages or learning objects are returned via search engines with no particular order. Regardless of the possibility that related, a client pushes ahead/backward in the material to make sense of the page to be perused first as clients more often than not have next to zero involvement in that domain. Despite the fact that a client may have domain instinct they are still to be connected. A learning way development approach based on changed TF-IDF, ATF-IDF and Formal Concept Analysis algorithms was proposed by [9]. The new approach initially built as Concept Lattice with keywords extracted by ATF-IDF from documents to guarantee a relationship hierarchy between keywords spoke to concepts. At that point FCA was utilized to compute intra-document relationships to settle on a right learning way.

Data classification for cross domains were researched and is a fundamental method to recognize one from another, as it needs to realize what has a place with which group. It can induce concealed dataset with obscure class through structural similitude analysis of a dataset with known classes. Classification results reliability is significant. The higher the produced classification results accuracy, the better the classifier. They routinely try to enhance classification accuracy through either existing methods or through growing new ones. Different systems are utilized to enhance classification accuracy performance. While most methods attempt to enhance classifier procedures accuracy, [10] decreased dataset features number by picking just relevant features before giving over dataset to classifier. Subsequently inspiring requirement for methods equipped for selecting relevant features with brought down information loss. The point is to diminish classifier workload utilizing feature selection. The review uncovers that classification with feature selection delivered noteworthy results with accuracy.

Feature selection has picked up significance because of its commitment to spare classification cost concerning time/computation load. Searching for crucial features, a feature search method is through decision trees. The last is an intermediate feature space inducer to choose vital features. A few studies utilized decision tree as feature ranker with direct threshold measure in decision tree based features selection, while others remain decision trees yet utilize pruning which goes about as a threshold instrument in feature selection. [6],[11] proposed a threshold measure utilizing Manhattan Hierarchical Cluster distance for use in feature ranking to choose relevant features as part of feature selection system. Results were promising and can be further enhanced by including higher number of qualities test cases.

Feature selection lessens features number in applications where data has 100's/1000's of features. Present feature selection concentrates on finding relevant features. Feature relevance is lacking to guarantee effective high dimensional data feature selection. Feature redundancy was characterized/proposed to perform feature selection redundancy analysis. Another framework decoupling relevance analysis and redundancy analysis was proposed. A correlation-based method for relevance/redundancy analysis was created and examined its proficiency/effectiveness compared to agent methods.

Chief component analysis (PCA) is the pillar of data analysis - a black box utilized and normally inadequately caught on. [11] Dispelled this myth as the manuscript planned to construct a strong instinct for how/why PCA functions. It solidified this knowledge by inferring the science behind PCA from basic instincts It was felt that by tending to all viewpoints, all readers would have an enhanced PCA understanding furthermore the when, how and why of this current procedure's application.

In this paper, it is proposed to compute the inverse document frequency and select features utilizing proposed feature selection. The effectiveness of the features therefore chose is evaluated utilizing LVQ classifier. It is proposed to extract the feature set from RatingSystem.com data set.

2. METHODOLOGY

2.1. Rating System.com Database

The Rating System.com is helping businesses worldwide to enhance the online shopping knowledge and interface with customers through the force of the product ratings, reviews, customer Q&A and social networking. It has an extensive database with relevant and comprehensive information on customer reviews. It started as a shell scripts set and data files. RatingSystem.com utilizes two methods to add information to a database: Web forms and email forms. Information from accommodation strategies shows that, it is easier to utilize web forms as opposed to email format, if just expansion to information is an overhaul. On the off chance that new information is to be submitted, clients ask for or get format layouts from RatingSystem.com through email. The proposed information must be formatted by and accepted.

2.2. Proposed Feature Selection Based on Decision Trees

Decision trees are prevalent methods for inductive inference. They are robust to noisy data and learn disjunctive expressions. A decision tree is a k-array tree in which each inside node indicates a test on a few characteristics from input feature set speaking to data. Every branch from a node relates to conceivable feature values determined at that node. Also, every test results in branches, speaking to changed test outcomes. The decision tree induction fundamental algorithm is a greedy algorithm building decision trees in a top-down recursive separation and-vanquish way.

The algorithm starts with tuples in the training set, selecting best characteristic yielding maximum information for classification. It creates a test node for this and after that a top down decision trees induction partitions current tuples set by test quality values. Classifier generation stops when all subset tuples fit in with the same class or on the off chance that it is not qualified to continue with extra partition to further subsets, *i.e.* in the event that more quality tests yield information for classification alone underneath a pre-indicated threshold. In this paper, it is proposed to construct the threshold measure based with respect to information addition and Manhattan progressive cluster.

In the proposed feature selection, a Decision tree induction chooses relevant features. Decision tree induction is the learning of decision tree classifiers building tree structure where each inward node (no leaf node) signifies property test. Every branch speaks to test outcome and every outside node (leaf node) indicates class prediction. At each node, the algorithm chooses best partition data credit to individual classes. The best credit to partitioning is chosen by trait selection with Information pick up. Quality with most elevated information pick up parts the trait. Information addition of the characteristic is found by

$$\text{info(D)} \quad = \quad -\sum_{i=1}^{m} p_i \log_2(p)$$

Where p_i is the likelihood, that self-assertive vector in D fits in with class c_i. A log capacity to base 2 is utilized, as information is encoded in bits. Data (D) is simply normal information amount required to distinguish vector D class name. The information addition is utilized to rank the features and the positioned features are dealt with as features in various leveled clusters. The proposed Manhattan distance for n number of clusters is given as takes after:

$$\text{MDist} \quad = \quad \sum_{i=1}^{n}(a_i - b_i)$$

A cubic polynomial equation is inferred utilizing the Manhattan values and the threshold measure is determined from the slope of the polynomial equation.

2.3. Learning Vector Quantization (LVQ)

Learning Vector Quantization (LVQ) is a local classification algorithm, where classification boundaries are locally approximated, the distinction being that as opposed to utilizing all training dataset focuses, LVQ utilizes just a prototype vectors set. This guarantees effective classification as vectors number requiring putting away or comparing is diminished incredibly. Furthermore, a deliberately picked prototype set additionally build noise issues in the classification accuracy [6].

LVQ is an algorithm that learns suitable prototype positions utilized classification and is characterized by P prototypes set $\{(m_j, c_j), j = 1\ldots \text{P}\}$, where mj is a K-dimensional vector in feature space, and c_j its class mark. The prototypes number is bigger than classes number. Hence, every class is spoken to by more than one prototype. Given an unlabeled data point x_u, its class name *yu* is determined as class c_q of closest prototype m_q

$$y_u = c_q, q = \arg\min_j d(x_u, m_j)$$

Where d is Euclidean distance. Other distance measures are utilized relying upon the issue.

3. RESULTS AND DISCUSSION

Features are extracted utilizing IDF from the Rating System.com data. The PCA and the proposed feature selection method was utilized to decrease the features. Table 1 demonstrates the classification accuracy got from LVQ and compared with Naïve Bayes classifier and Classification and Regression Tree (CART).

Table 1

Classification Accuracy

Technique used	*Classification Accuracy*
Naïve Bayes with LVQ	70
CART with proposed feature extraction	60.75
Naïve Bayes with proposed feature extraction	70.5
Naïve Bayes with proposed feature extraction	74.75

It can be seen from table 1, the classification accuracy got through Naïve Bayes with LVQ is superior to anything Naïve Bayes with PCA by around 5%. Figure 1 demonstrates the Root Mean Squared Error (RMSE). It can be seen that the accuracy and review low for the three classifiers.

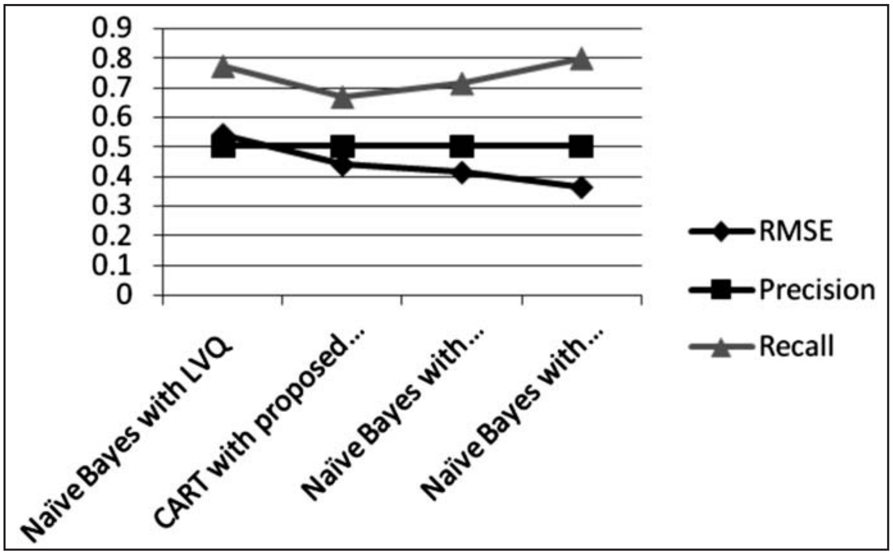

Figure 1: Root Mean Squared Error, Precision and Recall

4. CONCLUSION

Rapid advances in computer based high-throughput technique provided unparalleled chances for humans to expand production, services, communications, and research productions. Meanwhile, immense high-dimensional data quantities accumulate challenging state-of-the-art data mining techniques. Feature selection is needed for successful data mining applications, as they lower data dimensionality removing irrelevant features.In this paper, a feature selection for Sentiment Analysis using decision tree is proposed. LVQ type learning models constitute popular learning algorithms due to their simple learning rule, their intuitive formulation of a classifier by means of prototypical locations in the data space, and their efficient applicability to any given number of classes. Review features obtained from RatingSystem.com was extracted using inverse document frequency and the importance of the word found. The classification accuracy obtained by LVQ was 75%. However it was observed that the precision for positive opinions was quite low. This phenomenon was observed not only on LVQ but with Naïve Bayes classifier too.

5. REFERENCES

1. Vinodhini, G., & Chandrasekaran, R. M. (2012). Sentiment Analysis and Opinion Mining: A Survey. International Journal, 2(6).
2. Zhao, Z., Morstatter, F., Sharma, S., Alelyani, S., Anand, A., & Liu, H. (2010). Advancing feature selection research. ASU Feature Selection Repository.
3. El-Halees, A., & Gaza, P. (2011). Mining Feature-opinion in Educational Data for Course Improvement. International Journal of New Computer Architectures and their Applications (IJNCAA), 1(4), 1076-1085.
4. Omar, N., Jusoh, F., Ibrahim, R., & Othman, M. S. (2013). Review of Feature Selection for Solving Classification Problems. JISRI, 3.
5. Yacob, Y. M., Sakim, H. M., & Isa, N. M. (2012). Decision tree-based feature ranking using manhattan hierarchical cluster criterion. International Journal of Engineering and Physical Sciences.
6. Jeevanandam J & Dr. S. Koteeswaran, Decision Tree Based Feature Selection and Multilayer Perceptron for Sentiment Analysis, ARPN Journal of Engineering and Applied Sciences, ISSN 1819-6608, 2015, VOL. 10, NO. 14, 5883 – 5894.
7. Chen, J., Liu, Y., Zhang, G., Cai, Y., Wang, T., & Min, H. Sentiment Analysis for Cantonese Opinion Mining, Emerging Intelligent Data and Web Technologies (EIDWT), 2013 Fourth International Conference (pp. 496-500). IEEE.
8. Security in Wireless Sensor Networks: Key Management Module in EECBKM”Presented in International Conference on World Congress on Computing and Communication Technologies on Feb 27- & 28 and 1st march 2014, on St.Joseph college,Trichy
9. Sindhura, V., and Sandeep, Y. (2015, March). Medical data Opinion retrieval on Twitter streaming data. In Electrical, Computer and Communication Technologies (ICECCT), 2015 IEEE International Conference on (pp. 1-6). IEEE.
10. Weitzel, L., Aguiar, R. F., Rodriguez, W. F., and Heringer, M. G. (2014, June). How do medical authorities express their sentiment in Twitter messages?.InInformation Systems and Technologies (CISTI), 2014 9th Iberian Conference on (pp. 1-6). IEEE.
11. Bing, L., and Chan, K. C. (2014, December). A Fuzzy Logic Approach for Opinion Mining on Large Scale Twitter Data. In Proceedings of the 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing (pp. 652-657). IEEE Computer Society.