Analysis Of Machine Learning Data For Web Documents Using Fuzzy Clusters

Ms. N. ANURADHA
Associate Professor
Dept. Of Information Technology
Swami Vivekananda Institute Of Technology,Sec-Bad,Telangana,India.

Ms. D.S.V JYOTHI
Associate Professor
Dept. Of Information Technology
Swami Vivekananda Institute Of Technology,Sec-Bad,Telangana,India.

Ms. J. SUNANDA
Associate Professor
Dept. Of Information Technology
Swami Vivekananda Institute Of Technology,Sec-Bad,Telangana,India.

Abstract – *Web document clustering is one of the indispensable technique to discover contextual returned web pages are more complex there exists complicated association within one web document and linking to other user queries are also vague and unconscious numerous document clustering methods have been proposed some are based on probabilistic models equipped with distance similarity measures. Fuzzy linguistic space with a fuzzy clustering algorithm to discover the contextual meaning in the web documents that extract features of web documents using conditional random field methods and builds a fuzzy linguistic association of features. the feature organize a hierarchy of connected semantic complex called CONCEPTS wherein fuzzy linguistic measure is applied to evaluate the relevance of document belonging to a domain and difference between the other domains depending on the fuzzy linguistic measures.*

Keywords – Data mining, Fuzzy Logic, Neural Networks, Machine Learning.

Introduction I

Data mining is the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analysis offered by data mining move beyond the analysis of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour database for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Neural networks are of particular interest because they offer a means of efficiently modeling large and complex problems in which there may be hundreds of predictor variables that have many interactions.(Actual biological neural networks are incomparably more complex.) Neural nets may be used in classification problems (where the output is a categorical variable)
or for regressions (where the output variable is continuous).

A neural network (Figure 4) starts with an *input layer*, where each node corresponds to a predictor variable. These input nodes are connected to a number of nodes in a hidden layer. Each input node is connected to every node in the hidden layer. The nodes in the hidden layer may be connected to nodes in another hidden layer, or to an output layer. The output layer consists of one or more response variables.

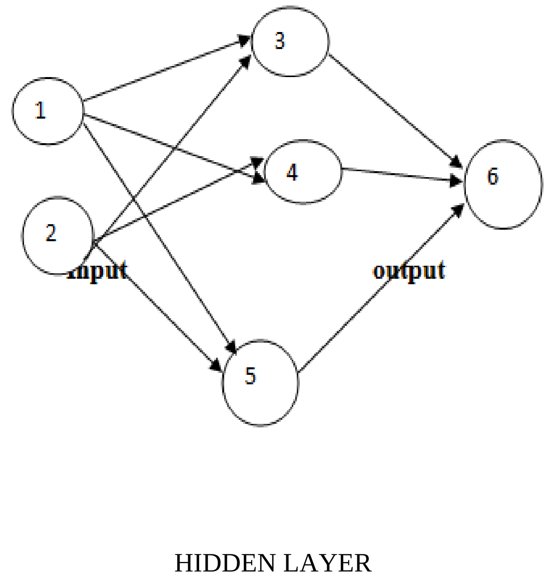


Figure 1 Neural network with one hidden layer

After the input layer, each node takes in a set of inputs, multiplies them by a connection weight Wxy (e.g., the weight from node 1 to 3 is W13 — see Figure 5), adds them together, applies a function (called the activation or squashing function) to them, and passes the output to the node(s) in the next layer. For example, the value passed from node 4 to node 6 is:

Activation function applied to ([W14 * value of node 1] + [W24 * value of node 2])

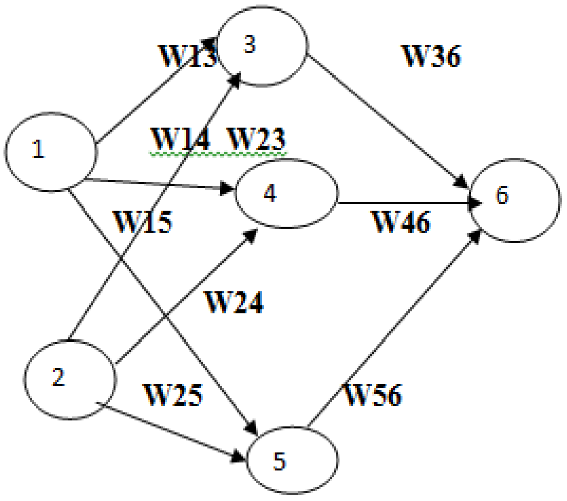


Figure 2 Wxy is the weight from node x to node y.

Each node may be viewed as a predictor variable (nodes 1 and 2 in this example) or as a combination of predictor variables (nodes 3 through 6). Node 6 is a non-linear combination of the values of nodes 1 and 2, because of the activation function on the summed values at the hidden nodes. In fact, if there is a linear activation function but no hidden layer, neural nets are equivalent to a linear regression; and with certain non-linear activation functions, neural nets are equivalent to logistic regression. The connection weights (W's) are the unknown parameters which are estimated by a training method. Originally, the most common training method was back propagation; newer methods include conjugate gradient, quasi-Newton, Leven berg-Marquardt, and genetic algorithms. Each training method has a set of parameters that control various aspects of training such as avoiding local optima or adjusting the speed of conversion. The architecture (or topology) of a neural network is the number of nodes and hidden layers, and how they are connected. In designing a neural network, either the user or the software must choose the number of hidden nodes and hidden layers, the activation function, and limits on the weights. One of the most common types of neural network is the feed-forward back propagation network for simplicity of discussion, we will assume a single hidden layer. Back propagation training is simply a version of gradient descent, a type of algorithm that tries to reduce a target value (error, in the case of neural nets) at each step. The algorithm proceeds as follows. Feed forward*:* The value of the output node is calculated based on the input node values and a set of initial weights. The values from the input nodes are combined in the hidden layers, and the values of those nodes are combined to calculate the output value.

Back propagation*:* The error in the output is computed by finding the difference between the calculated output and the desired output (i.e., the actual values found in the training set). Next, the error from the output is assigned to the hidden layer nodes proportionally to their weights. This permits an error to be computed for every output node and hidden node in the network. Finally, the error at each of the hidden and output nodes is used by the algorithm to adjust the weight coming into that node to reduce the error. The training set will be used repeatedly, until the error no longer decreases. At that point the neural net is considered to be trained to find the pattern in the test set because so many parameters may exist in the hidden layers, a neural net with enough hidden nodes will always eventually fit the training set if left to run long enough. But how well it will do on other data? To avoid an over fitted neural network which will only work well on the training data, you must know when to stop training. Some implementations will evaluate the neural net against the test data periodically during training. As long as the error rate on the test set is decreasing, training will continue. If the error rate on the test data goes up, even though the error rate on the training data is still decreasing, then the neural net may be over fitting the data.

Fuzzy logic is logic system for reasoning the approximate fuzzy sets that give universal set X in order to define fuzzy set A on X, where it define a membership A:X->[0,1] that maps element x of X into real numbers in [0,1] A(x) is interpreted as the degree to which x belongs to the fuzzy set A sometimes we call it as {x, A(x))|x € X}. the set theory contains certain element either belongs to a set or fuzzy set theory permits the gradual assessment of the membership of elements in relation to a set, consider U be universe of discourse representing a collection of objects denoted generically by u, fuzzy set A in a universe of discourse U is characterized by a membership function μ_A (u) = 1 is definitely member of A. A linguistic variable such as age may have a value such as young or its antonym old the great utility of linguistic variables can be modified via linguistic hedges applied to primary with certain functions.

SECTION II

2.Related Work: Fuzzy techniques for data mining presents theoretical common applications decomposed into two elements the notion of similarity and the fuzzy machine learning techniques that are applied in the described applications generally comparison measures are used at all levels of the data mining and information retrieval tasks at the lowest level for matching between a query to database and the elements it contains for the extraction of relevant data. Fuzzy machine learning that use the previous similarity measure is an important way to extract knowledge from sets of cases especially in large scale databases, we consider fuzzy machine learning methods that are used in the application aside other techniques as for instance fuzzy case based reasoning or fuzzy association rules that belongs to supervised learning framework that each data point is associated with a category unsupervised learning no priori decomposition of the dataset into categories is available. Fuzzy decision trees are particularly for data mining and information retrieval because enable the user to take into account imprecise description of heterogeneous values that appreciated for their interpretability provide a linguistic description of the relations between the cases and decision to make a class to assign rules for the user to interact with the system or the expert to understand confirm his own knowledge in the robustness descriptions does not drastically change the decision or the class associated with a case which guarantees a resistance to measurement errors and avoids values of the descriptions. Fuzzy protypes constitute approach to the characterization of data that interpretable of data sets to help better element to represent a group of data to summarize underline characteristic feature from statistic point of view as the data mean to median more complex representatives can also be used as the most typical value for instance, the prototype underlines the common feature of the category members but also the distinctive features as opposed to other categories underlining the specificity of the group some members of the group are more representative than others the typicality of a point depends both on its resemblance to other members of the group and on its dissimilarity to members of the other groups.

Supervised learning generates a function that maps inputs to desired outputs. For example, in a classification problem, the learner approximates a function mapping a vector into classes by looking at input-output examples of the function.

Classification is a data mining (machine learning) technique used to predict group membership for data instances. For example, you may wish to use classification to predict whether the weather on a particular day will be “sunny”, “rainy” or “cloudy”. Popular classification techniques include decision trees and neural networks.

Unsupervised learning models a set of inputs: labeled examples are not available.

Clustering is the method by which like records are grouped together. Usually this is done to give the end user a high level view of what is going on in the database. Clustering is sometimes used to mean segmentation – which most marketing people will tell you is useful for coming up with a birds eye view of the business. Two of these clustering systems are the PRIZM system from Claritas corporation and Micro Vision from Equifax corporation. These companies have grouped the population by demographic information into segments that they believe are useful for direct marketing and sales. To build these groupings they use information such as income, age, occupation, housing and race collect in the US Census.

SECTION III

3. Problem Definition: Nowadays web documents are more complex exist a complicated association within one web document and linking to the others, the high interactions between terms in documents. Techniques, such as TFIDF, have been proposed to deal with some of these problems. The TFIDF value is the weight of features in each document. While considering relevant documents to a search query, if the TFIDF value of a feature is large, it will pull more weight than features with lesser TFIDF values. The TFIDF value is obtained from two functions tf and idf, where tf (Term frequency) that appears in a document, and idf (Inverse document frequency), where document frequency is the number of documents that contain the feature.

3.1. Web Document Word Search: The service takes a term or phrase, and returns the different field of uploaded files that these could refer to. By default, it will treat the entire query as one term, but it can be made to break it down into its components. For each component term, the service will list the different filed (or concepts) that it could refer to, in order of prior probability so that the most obvious senses are listed first. For queries that contain multiple terms, the senses of each term will be compared against each other to disambiguate them. This provides the weight attribute, which is larger for senses that are likely to be the correct interpretation of the query.

Search engine is a class of software designed to search World Wide Web document for keywords entered by a searcher retrieves a list of documents containing the similarities in the basic way retrieve the index documents usually dispatches program to find from the web.

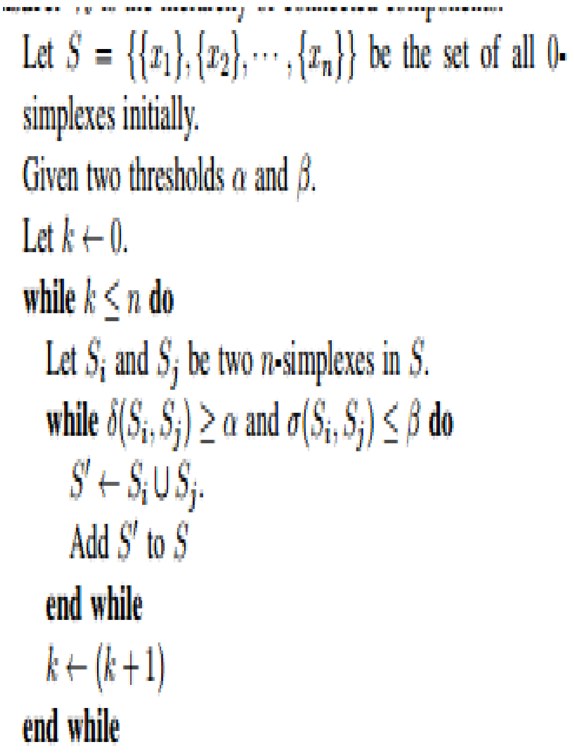
Fuzzy queries most easily performed through arguments to the match query type which has an extra A, should still turn up the vacuum product fuzziness should only be used with values of 1 and 2 a maximum of two edits between the query and a term in a document is allowed efficiently and are not processed.

3.2. Disambiguation in Web: Disambiguation cross-references each of these anchors with one pertinent sense drawn from the Page catalog, This phase takes inspiration from but extends their approaches to work accurately and on-the-fly over short texts. Aim for the collective agreement among all senses associated to the anchors detected in the input text and we take advantage of the un-ambiguous anchors (if any) to boost the selection of these senses for the ambiguous anchors. However, unlike these approaches, we propose new disambiguation scores that are much simpler, and thus faster to be computed, and take into account the sparseness of the anchors and the possible lack of un-ambiguous anchors in short texts.

3.2. Parsing Technique: Parsing detects the anchors in the input text by searching for multi-word sequences in the upload file field category. Tagme receives a short text in input, tokenizes it, and then detects the anchors by querying the Anchor upload file field category for sequences of words.

3.3. Fuzzy Hierarchical Alogirthm: We documents are clustered based on maximal simples of any dimensions that web documents clustered by CONCEPTS contains common lower dimensional faces which is a consequences of Apriori property sense the methodology provides a soft approach wherein we allow lower dimension to overlap with CONCEPTS existing across many clusters. Input is collection of documents performs feature extractions using CRF methods to generate entities, Calculate the fuzzy linguistic value of every named entity to every named category. Perform hierarchical aggregation clustering in order to generate the semantic hierarchy from the set of coincident entities.

V = A set of $\{x_1, x_2 \ldots x_n\}$be the vertex set of all reserved named entities generated from W associated with the categories in a collection of documents, H is the hierarchy of connected components.



```
Let $S = \{\{x_1\},\{x_2\},\cdots,\{x_n\}\}$ be the set of all 0-simplexes initially.
Given two thresholds $\alpha$ and $\beta$.
Let $k \leftarrow 0$.
while $k \le n$ do
  Let $S_i$ and $S_j$ be two $n$-simplexes in $S$.
  while $\delta(S_i, S_j) \ge \alpha$ and $\sigma(S_i, S_j) \le \beta$ do
    $S' \leftarrow S_i \cup S_j$.
    Add $S'$ to $S$
  end while
  $k \leftarrow (k+1)$
end while
```

The alogirthm starts the set of all 0-simplexes that is single named entities aggregate web documents into several categories to be the primitive concepts if the web document contains a primitive concepts means that web document highly equated to such concept with reference to Apriori property all the sub clusters in the concept are also contained in the web document classify generally web consist of primitive concepts into multi-categories.

SECTION IV

4. Analysis Fuzzy Logic in Web Document: The documents provide imprecise information; the use of fuzzy set theory is advisable. Fuzzy c-means and fuzzy hierarchical clustering algorithms were deployed for document clustering. Fuzzy c-means and fuzzy hierarchical clustering need prior knowledge about ‘number of clusters’ and ‘initial cluster cancroids’,’ which are considered as serious drawbacks of these approaches. To address these drawbacks, ant-based fuzzy clustering algorithms and fuzzy k- means clustering algorithms were proposed that can deal with unknown number of clusters.

The system extracts features from the web documents using conditional random field methods and builds a fuzzy linguistic topological space based on the associations of features. The associations of co-occurring features organize a hierarchy of connected semantic complexes called ‘CONCEPTS,’ wherein a fuzzy linguistic measure is applied on each complex to evaluate (1) the relevance of a document belonging to a topic, and (2) the difference between the other topics. The general framework of our clustering method consists of two phases. The first phase, *feature extraction*, is to extract key named entities from a collection of “indexed” documents; the second phrase, *fuzzy clustering*, is to determine relations between features and identify their linguistic categories.

CONCLUSION V

A single term is not able to identify a latent concept in a document, for instance, the term “Network” associated with the term “Computer,” “Traffic,” or “Neural” denotes different concepts. A group of solid co-occurring named entities can clearly define a CONCEPT. The semantic hierarchy generated from frequently co-occurring named entities of a given collection of web documents, form a simplified complex. The complex can be decomposed into connected components at various levels (in various levels of skeletons). We believe each such connected component properly identify a concept in a collection of web documents.

Reference

[1] Keyword and search engines statistics. http://www.keyworddiscovery.com/keyword-stats.html?date=2013-01-01 (2013)

[2] Atallah, M.J., Frikken, K.B.: Securely outsourcing linear algebra computations. In: Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security, pp. 48–59. ACM (2010)

[3] Atallah, M.J., Li, J.: Secure outsourcing of sequence comparisons. International Journal of Information Security **4**(4), 277–287 (2005)

[4] Attrapadung, N., Libert, B.: Functional encryption for inner product: Achieving constant-size ciphertexts with adaptive security or support for negation. In: Public Key Cryptography–PKC 2010, pp. 384–402. Springer (2010)

[5] Azab, A.M., Ning, P., Zhang, X.: Sice: a hardware-level strongly isolated computing environment for x86 multi-core platforms. In: Proceedings of the 18th ACM conference on Computer and communications security, pp. 375–388. ACM (2011)

[6] Bao, F., Deng, R.H., Ding, X., Yang, Y.: Private query on encrypted data in multi-user settings. In: Information Security Practice and Experience, pp. 71–85. Springer (2008)
[7] Boneh, D., Di Crescenzo, G., Ostrovsky, R., Persiano, G.: Public key encryption with keyword

[8] H. L. Larsen, "An approach to flexible information access systems using soft computing," in Proc. of the 32nd Annual Hawaii International Conference on System Sciences, Hawaii, 1999, p. 231.

[9] W. B. Frakes and R. Baeza-Yates, Information Retrieval Data Structures and Algorithms. Englewood Cliffs, NJ: Prentice Hall, 1992.

[10] S. Park, D. U. An, B. R. Cha, and C. W. Kim, "Document clustering with cluster refinement and non-negative matrix factorization," in Proceedings of the 16th International Conference on Neural Information Processing, Bangkok, Thailand, 2009, pp. 281–288.

[11] T. Kohonen, Self-Organization Maps. Berlin Heidelberg: SpringerVerlag, 1995. [12] J. MacQueen, "Isome methods for classification and analysis of multivariate observations," in Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1. Berkeley, CA: University of California Press, 1967, pp. 281–297.

gent