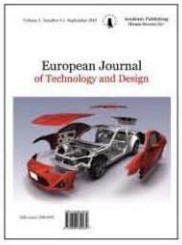
ISSN: 2310-0133
Founder: Academic Publishing House *Researcher*
DOI: 10.13187/issn.2310-0133
Has been issued since 2013.

European Journal of Technology and Design



A Brief Review of Main SVM-based Ranking Models

[1] Igor Yu. Botian
[2] Lev V. Utkin

[1-2] Saint Petersburg State Forest Technical University, Russian Federation
Department of Control, Automation, and System Analysis
[1] E-mail: igor.botian@gmail.com
[2] E-mail: lev.utkin@mail.ru

Abstract. The paper deals with a critical review of SVM-based ranking models which are used to solve a certain set of machine learning problems. A short description, advantages, and disadvantages are provided for each model, as well as general comparison of them. A problem connected with getting a training data for these models is also raised. Applications of the models are given, as well as future research directions.

Keywords: Machine learning, ranking, support vector machine, training data, comparison judgments.

1. Introduction

Nowadays amount of information which a person has to deal with every day increases steadily. In this connection, a matter of value of the information being look-up by the person or incoming for him is raised inevitably. In other words, a problem of information relevance is emerged. Under conditions of broad proliferation of computer systems, one of the main means of solving such the problem is so-called search engines.

Depending on their nature, they provide different information to user. For instance, web search engines provide relevant results of user queries. At that, depending on a type of the query, user can get a list of web-links, the result of an entered mathematical expression, forecast information, and so on, as results of his query. There are engines providing an answer for a user question asked in natural language (Apple Siri, Google Now, Microsoft Cortana). Services like Shazaam searches for name of a song fragment recorded by the user. Finally, there are many online advertising services. The content of the advertising banner exposure to user is based on his search history, visited web-sites, his geolocation, etc.

Each search engine is based on the principle of data ranking. Thanks to this fact, depending on the query the same data can correspond to different relevance degrees. Information about relevance one or another data is taken from so-called training data. Values of all elements belonging to the training data are given by experts.

One of the most used approaches to learning to rank is an approach based on *support vector machine* (SVM) apparatus [1]. Based on its features, ranking problem reduces to data classification problem. It can be said that it may be treated as a general tool used for such the set of problems.

Classification problems are specific enough thus depending on application domain they may differ a lot. For example, a standard form of training data used in classification in the context of search engines cannot be obtained at all [2]. In these systems, a priori information which can be used in many other cases is represented in a form of some relative assessments. As a consequence of this fact, there is a need of development of appropriate approaches oriented on such the data processing.

2. SVM-based ranking models

By definition, the learning to rank problem is referred to a type of *supervised machine learning*. Such the learning consists of ranking model automatic selection on a given training data set. One of the ranking models is a model based on support vector machine method [1]. From mathematical point of view, the method consists in mapping of vectors representing input data into space with greater dimensional and further searching of a separating hyperplane with maximum gap in this space. The main advantage of this method is efficient realization non-linear separating functions in the context of the classification problem.

SVM-based models, as well as any ranking models, can be divided into three categories depending on the approach they based on [2]: pointwise, pairwise, or listwise. The main difference consists in what form input and output data have, as well as loss function.

1. **Pointwise approach.** It is expected that each input-output data pair confronts with a certain numerical assessment. In this case the learning to rank problem reduces to regression building: a numerical assessment is needed to predict for each single pair. In the bounds of this approach many existing machine learning regression-based algorithms can be applied directly.
2. **Pairwise approach.** While the approach is used, learning to rank problem reduces to building a binary classifier dealing with two input objects corresponding to the same query. It is needed to determine which of them is more relevant.
3. **Listwise approach**. The approach consists in building a model dealing with a set of all input objects corresponding to the same query. As a result it returns a list of these objects but ranked.

The main SVM-based models corresponding to these approaches are SVM-based pointwise model [3], RankSVM [4], and SVM^{map} [5]. Review of each model is given below in further sections.

3. SVM-based pointwise model

The model is based on pointwise approach [3]. A principle of work consists in reduction of ranking problem to classification problem. In this case the classification is regarded to a variation of *supervised learning problem*, in which all output values are discrete. At that, an output value corresponds to a relevance degree of a given object. In this connection, the value relates to a certain object category. A rigorous mathematical definition of the algorithm and the corresponding optimization problem which implements this model are given in [3].

There is a disadvantage connected with SVM-based pointwise model. By its nature, ranking is more relative order prediction rather than a precise relevance degree of a given object. Thus while the model is applied, a relative order between objects can not normally take a part in the learning process by the reason of its absence.

4. RankSVM

The models are based on pairwise approach [4]. And the ranking problem in this case reduces to a binary classification problem. In comparison to SVM-based pointwise model, an output value is a relative order between objects, no an absolute value which characterizes a relevance degree of them. Thus, RankSVM does not have the disadvantage which is belonged to SVM-based pointwise model.

In its turn, RankSVM has a number of disadvantages. At first, by its nature, a result of binary classification is a list of objects which contains no information about difference in relevance degree between each two next objects. At second, RankSVM is more sensitive to imprecise data from the training set. It affects negatively on quality of obtained results. At last, dependence of a number of pairs on a number of objects is mathematically quadratic that negatively impacts on processing performance in case of a huge data.

Initially, the method was developed to improve search engines efficiency. In simple words, it is easier to describe this model in terms of web-search. In whole, RankSVM is comprised of three steps on learning stage:

1. It calculates distances between two any vectors given on step 1.
2. It maps input elements between queries and web-pages visited by the user into a certain feature space.
3. It states an optimization problem similar to SVM classification and solves such the

problem by means of a regular SVM solver.

A rigorous mathematical definition of the algorithm and the corresponding optimization problem which implements RankSVM are given in [4].

5. SVM^{map}

A principle of work of SVM^{map} is based on listwise approach [5]. As a whole, the model solves classification problem in more natural way rather than SVM-based pointwise model and RankSVM. It can be explained that these two models do not take into account the fact that some objects (or pair of objects) do not correspond to the same input query.

In other words, SVM^{map} uses a whole set of objects corresponding to a query as one input object and a list of these input objects (or their relevance degrees) in ranked form. A rigorous mathematical definition of the algorithm and the corresponding optimization problem which implements SVM^{map} are given in [5].

In accordance to previously conducted researches [6], efficiency of listwise ranking algorithms is better in whole rather than pointwise and pairwise ones. On other hand, a more efficient learning algorithm is needed in order to make the listwise approach more practical. It also regards to SVM^{map} model.

6. SVM-based ranking models comparison

SVM-based pointwise model reduces ranking to regression, classification, and ordinal regression. In its turn, RankSVM reduces ranking to pairwise classification, whereas SVMmap reduces ranking to a brand-new problem and defines specific algorithms for this.

Advantage of the first two models consists in possibility of direct application of many existing various tools. The main disadvantage of them is that individual ranking features are not considered in definition of algorithms used by them. An issue related with SVM^{map} consists in difficulty of algorithms used by the model.

7. Problems connected with training data set

On the assumption of application of the models given above and taking into account of their advantages and disadvantage, there is a choice of an appropriate model needed in order to solve a certain practical problem. At that, to build an efficient model another important problem should be taken into account. It consists in preparation of a high qualitative training set. The process, in its turn, connects with a set of problems:

- At first, the fact should be taken into account that a training set is needed to have a great number of objects but at the same time have a cost of its getting as less as possible. A striking example which exposures this issue is a straightforward inquiry of a great number of respondents: an amount of data obtained is limited and the process itself is costly.
- At second, data obtained by experts is not always enough and even correct because there may be people subjectivity taken a place during inquiry process. Development and further application of the data selection strategy can be regarded as a solution for this issue. First examples of such the strategies are Depth-k pooling [7], MTC [8], and LETOR [9].

To reduce cost of a training data acquisition it is logical to approach to automation of this process. In this connection, there is a need to develop and apply a strategy which favors more well thought-out selection of training data elements. In extreme case, it may help to get rid of over-fitting issue. In its turn, when it is possible, analysis of actions made by the user can be applied to promote further adaptation of used algorithms.

The issues given above are not fully solved in the bounds of existing models. And not only SVM-based models but others also. At that, it should be noted that these problems are just main whereas the process of training data acquisition is conjugated with a number of additional problems [2].

8. Future research directions

Nowadays there many researches are conducted in the area of learning to rank problems. In this connection, new and new problems required further investigation are emerged. At that, it can be noted that they affect various aspects of the ranking problem.

As it was said before, ranking algorithms based on listwise approach are researched in less degree rather than algorithms based on pointwise and pairwise approaches. At that, it is noted a need to develop more advanced ranking models rather than existing ones [10].

A problem related to building a training data set takes a place. Generally, the dataset is comprised of nonrandom objects selected manually. Each of them affects on results of the learning in some degree, whereas real data may have a rather big percentage of objects which correspond to a small relevance degree. Solution of this problem is connected with well-thought selection of objects belonging to the training data set.

There is another problem related to a training data set [2]. By its nature a training data set built basing on experts' assessments does not confront by its scale with amount of data needed for rigorous learning. At that, possibility to adapt existing algorithms to new incoming data is not taken into consideration. Such the issue can be solved by means of so-called self-learning of ranking algorithms, when processed data takes a part in the training set in a certain degree.

As of ranking algorithms, there are several problems connected with them. At this time a few attention is paid to scalability on development of the ranking algorithms [11]. The possibility to parallel the algorithm may considerably decrease its execution time in case of a huge amount of data. It is especially important in real-time systems, where a trade-off between results accuracy and algorithms time consumption is emerged inevitably [12]. In its turn, a problem connected with flexible selection of ranking features may be also referred to this category [13]. The training data set directly influences on results of data processing. At this moment it is built on the assumption of subjective experts' criterion. But on the other hand it might change at runtime to meet optimization needs. At that, the algorithm itself should be resistant to changes in the feature set and not to depend on it [13]. To some extent, it is also connected with self-learning of the ranking model.

At last, a ranking problem is emerged on conditions that the given information is imprecise or incomplete. In this case a priori relative assessments are group, or in other words interval. It should be noted that at this moment such the problem is not researched enough. Authors of this paper are working on this direction.

9. Conclusions

Aim of this work was to give a short critical review of existing ranking models based on the support vector machine apparatus. The following models were considered: SVM-based pointwise model, RankSVM, SVM^{map}. They are based on fundamentally different ways of input data processing. Main advantages and disadvantages were provided for each model. Comparison of these SVM-based models was given.

Along with SVM-based models review, existing problems connected with building a training data set were enumerated. Implementation of each model is based on this data set. Future research directions were given. From the authors point of view, one of the most challenging open issue deals with modification of existing ranking algorithms is their adaptation to imprecise and limited relative assessments.

10. References

1. Herbrich, R., Obermayer, K., Graepel, T.: Large margin rank boundaries for ordinal regression. In: Advances in Large Margin Classifiers, pp. 115–132 (2000).
2. T.-Y. Liu: Learning to Rank for Information Retrieval, Springer (2011).
3. Nallapati, R.: Discriminative models for information retrieval. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004), pp. 64–71 (2004).
4. Joachims, T.: Optimizing search engines using clickthrough data. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2002), pp. 133–142 (2002).
5. Yue, Y., Finley, T., Radlinski, F., Joachims, T.: A support vector method for optimizing average precision. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007), pp. 271–278 (2007).
6. Xia, F., Liu, T.Y., Wang, J., Zhang, W., Li, H.: Listwise approach to learning to rank—theorem and algorithm. In: Proceedings of the 25th International Conference on Machine Learning (ICML 2008), pp. 1192–1199 (2008).

7. Aslam, J.A., Kanoulas, E., Pavlu, V., Savev, S., Yilmaz, E.: Document selection methodologies for efficient and effective learning-to-rank. In: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009), pp. 468–475 (2009).

8. Carterette, B., Allan, J., Sitaraman, R.: Minimal test collections for retrieval evaluation. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006), pp. 268–275 (2006).

9. Liu, T.Y., Xu, J., Qin, T., Xiong, W.Y., Li, H.: LETOR: benchmark dataset for research on learning to rank for information retrieval. In: SIGIR 2007 Workshop on Learning to Rank for Information Retrieval (LR4IR 2007) pp. 3-10 (2007).

10. Qin, T., Liu, T.Y., Zhang, X.D., Wang, D., Li, H.: Learning to rank relational objects and its application to web search. In: Proceedings of the 17th International Conference on World Wide Web (WWW 2008), pp. 407–416 (2008).

11. Chang, E., Zhu, K., Wang, H., Bai, H., Li, J., Qiu, Z., Cui, H.: Parallelizing support vector machines on distributed computers. In: Platt, J., Koller, D., Singer, Y., Roweis, S. (eds.) Advances in Neural Information Processing Systems 20 (NIPS 2007), pp. 257–264. MIT Press, Cambridge (2008).

12. Wang, L., Lin, J., Metzler, D.: Learning to efficiently rank. In: Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010), pp. 138–145 (2010).

13. Chang, Y., Liu, T.Y.: Future directions in learning to rank. JMLR: Workshop and Conference Proceedings 14, pp. 91–100 (2011).