It is well known that in order to handle the overfitting of machine learning algorithms one should control the search space complexity, for example the number of hidden neurons and layers in the neural net, the number of nodes and leafs in decision trees, etc. At the same time, the theoretical bounding of overfitting remains one of the most challenging open problems in computational learning theory. Starting from VC-theory (Vapnik and Chervonenkis, 1971), this problem was addressed via many different techniques. Recent works are mainly based on concentration and deviation inequalities by Talagrand, Rademacher sums, PAC-Bayesian bounds, and others tools.

In current work we follow the combinatorial approach to the problem of overfitting, where the search space is represented as a finite set of hypotheses with known behaviour on finite set of data points. For a given data point, each hypothesis either make a mistake, or classify it correctly. Therefore each hypothesis is represented as a binary vector of its errors, and whole search space - as a binary error matrix.

In this framework the overfitting is defined as the score of complete cross-validation procedure. We use all possible ways to independently split all data points into train and test sample of a fixed length. More precisely, for a given pair of train and test sample we select best hypothesis based on its error rate on the train sample, and calculate the deviation with its error rate on test sample. The probability of overfitting is defined as the fraction of test samples with significant deviation between test and train error rates.

It has been experimentally shown that there are two major factors --- splitting and connectivity, which reduce overfitting. The splitting property means that only a small part of hypotheses from a given search space have a low error rate and, as a result, a high chance to be produced by a learning algorithm. The connectivity property means that in most practical search spaces there are many hypotheses that are similar to each other in terms of Hamming distance between their error vectors.

Splitting and connectivity are formally defined in terms of splitting-connectivity graph (SC-graph) and splitting-connectivity profile (SC-profile). Each vertex of SC-graph corresponds to a single hypothesis, and two hypotheses are linked together if they differ on exactly one data point. SC-profile represents some properties of SC-graph, such as the amount of vertices with certain number of edges.

In this work we study some properties of SC-profiles for binary classification problem. We show that in this special case the connectivity is a purely geometrical concept, so that it only depends on the location of data points in the space, but not on their target classes. We randomize SC-profile over all possible assignments of target classes to the data points, and prove that average SC-profile is the product of two independent profiles - the splitting profile and the connectivity profile. We also prove theoretically that the average splitting profile is given by binomial distribution. We use experiments on model data to illustrate that real splitting profile actually follows binomial distribution when target classes are randomly assigned to the data, and study how the profile changes when the assignment of target classes follows some underlying pattern. Finally, the geometrical properties of SC-profiles will be demonstrated for the set of linear classifiers in three-dimensional space and for the set of conjunctive rules.